PERSONALISED TASTE PROFILING IN SHORT-TEXT MICROBLOGS

A THESIS SUBMITTED TO AUCKLAND UNIVERSITY OF TECHNOLOGY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Supervisors

Dr Muhammad Asif Naeem

Dr Farhaan Mirza

Associate Professor Russel Pears

30th March 2021

By

Herman Masindano Wandabwa

School of Engineering, Computer and Mathematical Sciences

Abstract

The objective of this thesis is to develop diverse and user-representative methods for taste profiling in short-text microblog users. The proposed methods are entirely based on the disseminated content, social network structure and their variations over time. Inferring user interests and subsequent formulation of taste profiles is pertinent in *personalizing content recommendations* for micro-blogging services as well as in *extraction of users with similarities in preferences*. The methods are broadly divided into two categories: i) short-text analytics methods (Part I, Chapter 3) and ii) user interest identification and quantification (taste profiling) over time (Part II, Chapters 4,5 and 6).

With the proposed method in Part I, it is possible to accurately extract knowledge from short texts, a usually difficult process due to the unconventional language on such platforms. As a case study, a semi-supervised modelling framework is proposed based on tweets metadata in extraction of better topical representations of short texts. In the findings, topical vectors from semantically relevant long texts made shorter and otherwise noisy texts more interpretable. The built models generated better results in terms of topical classifications compared to similar approaches.

The methods in Part II largely support the detection of user interests and subsequent modelling of taste profiles. As case studies, several approaches were proposed in identifying and quantifying short-text microblog users' interests. A neural network-based approach was proposed in the computation of user interests in a specific topic as part of the process to identify relevant users for follow-back feature in certain domains.

In addition, a soft clustering method was proposed to identify user interests in several topics and to certain levels. Lastly, the time dependency factor in interest decay and gain in such microblogs was modelled. This mirrored a conventional short-text microblogging platform where content is volatile based on for example the prevailing news at the time. Twitter was used as the testing platform for the proposed approaches mainly because of its popularity, API access ability as well as the temporal-dynamism of its overall network structure. This research is fundamental to services, content recommendations and audience measurement.

Contents

List of Tables

List of Figures

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the qualification of any other degree or diploma of a university or other institution of higher learning.

Signature of student

Acknowledgements

Firstly, I'm grateful to the Almighty God for the precious gift of life in my course of PhD studies. My research would not have taken shape were it not for the support of the supervisory team. I'm so grateful to the team especially to Dr Muhammad Asif Naeem who put his heart and soul into my research and made sure that we delivered as a unit. His persistence, hard work and support is immeasurable. I also extend my sincere gratitude to Dr Farhaan Mirza, my secondary supervisor. His ability in having this research translated to a measurable market level product pushed me to work hard on it. Special thanks to Associate Professor Russel Pears for always being there when we all needed support in clarification of fundamental machine learning concepts and general guidance in shaping up the research topic. I will forever be grateful.

My research would not have taken shape were it not for the support of Auckland University of Technology and the School of Engineering, Computer, and Mathematical Sciences. Your financial support, provision of a conducive research environment as well as general support for my well-being is worth appreciation. You positively contributed to this research and thesis as the end product.

I'm grateful to the people who left an indelible mark in my life. My former teachers especially *Prof. Franklin Wabwoba* for being the first person to introduce me to the first computer at *Friends School Kamusinga*. My friends who have been family while away from home, were the pillars when I needed a good laugh during difficult research times. Blessings to *Dennis Munene* and *Wandia, Richmond Ogutu, Peter Kipkenei, Steve Gichure, Susan Njeri* and *Amara*, *Joy Chemelil, Buka Abbas, George Opiyo* and *family, Kevin Nthiga, Peter Makolo* and *family, Bikash Pokhrel* and *family* and, *Dickson Lukuba* for always extending the brotherly hand.

Lastly, special appreciation goes to my entire family, and more specifically my parents in Kenya for the sacrifices they made for me to get an education. My late father's passion for an educated family inspired me to reach this academic milestone. My mother's fervent prayers every time I left home for school positively influenced me to work hard in this academic journey. Thank you for the support.

As for *Naomi Barongo*, and *Davyn*, you deserve a line of your own for your presence, encouragement and prayers. Thank you so much.

Dedication

*Dedicated to my parents, posthumously to the late **Mr Alfred Yuka Wandabwa**, and **Mrs Billiah Nanjala Wandabwa**, the pillar of our family. Your sacrifices for me to get an education, and the values that you instilled in me, were not in vain.*

Chapter 1

Introduction

This chapter provides an overall background of taste profiling concepts in short-text microblogs including the research problem and motivation for this research. Additionally, the research questions, objectives, overview of the research solution, contributions and the overall research roadmap in the discernment of user-representative profiles in short-text microblog. The flowchart in Figure 1.2 provides an overview of the entire research process from modelling short and noisy texts through to formulation of taste profiles over time.

1.1 Research Problem

Online content propagation in present time has been proliferated following the surge in citizen journalism. This is partly attributed to the increase in the number of devices such as mobile phones, as well as the emergence of many social microblogging platforms such as Twitter[1], Facebook [2], etc. Access to cheaper internet and mobile devices has in turn given rise to a vibrant web in which many people communicate, regardless of their geographical locations. Short-text microblogs such as Twitter have been instrumental

[1] https://twitter.com/
[2] https://www.facebook.com/

in the dissemination of near to real-time data in the form of text, videos, hyperlinks and images. Statistically, the number of disseminated tweets averages about 200 billion per year, which roughly translates to approximately 6,000 tweets per second and over 350,000 tweets sent per minute [3]. On the platform, users are able to re-share the disseminated tweets, in form of "retweets" as well as "comment" on and/or "like" the original tweet. The process of pursuing user representative knowledge from such short-text microblogs for better provision of third-party related services is known as *user taste profiling*.

Therefore, User-Generated Content (UGC) is pertinent in microblogs as the content is valuable in applications such as event detection (Sakaki, Okazaki & Matsuo, 2010) and discovery of new web content (Dong et al., 2010) among others. User interests are an important discovery in UGCs as they define the *user profiling* process to a large extent. User profiling is defined in several ways in previous studies as well as based on the nature of the application domain. Ouaftouh et al. (Ouaftouh, Zellou & Idri, 2015) and Zhou et al. (Zhou, Xu, Li, Josang & Cox, 2012) define user profiling as the identification of representative descriptors of a user. This consists of demographically descriptive information such as the user's name, country, level of education, age etc. Such information represents user preferences or interests. Alaoui et al. (Alaoui, Idrissi & Ajhoun, 2015) defined the user profiling as the process of gathering information that offer insight to a user's need and is able to forecast the user's future intention. The authors noted that such information depended on three factors: similarities, trace handling, and machine learning algorithms' predictions. More recently, Chen et al. (M. Chen, Ghorbani et al., 2019) described a *user profile* as a pattern that consists of user behavioral tendencies and preferences. In their study, they were of the opinion that mined knowledge about users behavioural patterns can be used to predict the user's future intentions.

[3] http://www.internetlivestats.com/twitter-statistics/

In microblogs, *user profiling* can refer to all or some of the user-specific information e.g., keywords in the profile, country, gender, etc., relevant to the user. User modeling with respect to user interests in Online Social Networks (OSNs) is subdivided into two categories based on the context with which they are learned. They are either: - (a) Latent user models, or (b) Transparent user models.

Latent models refer to profiles learned with task specificity at hand, e.g., a user model trained for tweets classification or recommendation based on the user's retweet history. It is specific to the set classification task and no other tasks (K. Chen et al., 2012), (Tang, Hu & Liu, 2013). The model is thus tied to a specific task with vectors of predefined dimensions. *Transparent models*, on the other hand, are unsupervised based on user activities, and thus are more interpretable and better suited for cold start scenarios (El-Arini, Paquet, Herbrich, Van Gael & Agüera y Arcas, 2012). For example, user models trained on the retweet history of many users can be used in the recommendation of tweets that are likely to be retweeted. The same can be said of third-party applications, e.g., news recommender systems that are based on social login functionality like Twitter's. In addition, formulation of user profiles in microblogs involves the interpretation of both *explicit* and *implicit* information. Implicit information has regard to what users disseminate, e.g., tweets. Such information usually needs clues from parts of the disseminated content to be comprehended. On the other hand, explicit information requires direct user input e.g. information declared by a tweeter upon sign up. Normally, such information is related to user specific interests. Generally, the main objective of any user profiling process is to acquire subject representative data that are bound to enhance the quality of user information access as well as to discern user intentions. This process is instrumental in tasks related to varied application areas, including third-party content recommendations.

At a high level, user taste profiles are either *static* or *dynamic* in their representations as long as the extracted profile is accurate in its representation (Farseev, Akbari,

Samborskii & Chua, 2016). In static profiles, data representation relies on the analysis of a user's predictable and otherwise static characteristics (Farseev et al., 2016), (Poo, Chng & Goh, 2003). Therefore, a static profile usually maintains long term descriptors of the user. For example, in microblogs like Twitter, user's profile metadata rarely changes despite variations in the disseminated content. Farseev et al., (Farseev, Nie, Akbari & Chua, 2015) made use of an ensemble model to construct user profiles from a multi-modal dataset from three geographical areas. From the findings, the fusion of multiple data sources improved the profiling process. In dynamic profiling, user descriptors change over time (Kanoje, Girase & Mukhopadhyay, 2015), (Yin, Cui, Chen, Hu & Zhou, 2015b). This is quite representative of short-text microblogs where the user disseminated information is used to model the user's taste profile. This information is inherently dynamic, and changes even in its semantic representation as the information decays over time. In environments where the velocity, volume and variety are key, such as in short-text microblogging platforms, then dynamism in the profiling process is paramount. Works by Akbari et al., (Akbari, Hu, Liqiang & Chua, 2016), in profiling tweeter's wellness, used aggregated information from tweeter's timelines with event categories, since the tweeter's social accounts implicitly reflected their preferences, habits and feelings. They utilised the graph Laplacian as a regulariser in discerning the inter-relatedness between what users disseminated to different events. Dynamic User and Word Embedding (DUWE) model and Streaming Keyword Diversification Model (SKDM) was used by Liang et al. (Liang, Zhang, Ren & Kanoulas, 2018a) in addressing the user profiling process in tweeters. The two models tracked the semantic representations users and terms over time and computing their similarities to present succinct profiles over time.

1.2 Research Gaps

The points in Section 1.1 highlight a few modelling approaches in extracting knowledge in Online Social Networks (OSNs). However, the below research gaps exist in the user taste profiling process, particularly in the domain of short and noisy texts. Details of how these gaps are addressed are in Sections 1.4 and 1.5.

1. **Underlying semantics** - Semantics i.e. the meaning or interpretation intended by a tweeter in short and noisy texts is one other problem that researchers have been working hard to address (Blei, Ng & Jordan, 2003a; W. X. Zhao, Jiang, He et al., 2011; W. X. Zhao, Jiang, Weng et al., 2011; Z. Liu, Huang, Zheng & Sun, 2010; Weng, Lim, Jiang & He, 2010a; Andrzejewski, Zhu & Craven, 2009; Z. Chen et al., 2013b; X. Wang, Wei, Liu, Zhou & Zhang, 2011). External knowledge in the form of ontologies has been used to augment the modelling of the disseminated content that is inherently short and has a sparse vocabulary. In dealing with this high dimensionality, concepts from such external ontologies, e.g., DBpedia/Wikipedia categories/subcategories have been incorporated in the modelling process (J. Wang, Zhao, He & Li, 2014), (Budak, Kannan, Agrawal & Pedersen, 2014), (Kapanipathi, Jain, Venkataramani & Sheth, 2014b), (Michelson & Macskassy, 2010). Inter-category relations can, therefore, be deduced to improve on the accuracy of the overall modelling process. However, the temporal nature of microblogs makes it difficult to achieve meaningful and time-bound results. To put this in perspective, it takes only a few minutes for a topic to "trend" on Twitter as long as it gains enough attention in a short span of time. For Wikipedia knowledge to be incorporated, a Wikipedia page and related categories on the same issue have to be created first, which often takes time or is sometimes totally ignored. Contributions in addressing this research gap are in Chapter 3.

2. **Poly-representation of interests** in short-text microblogs where, each topic of interest is considered a single concept makes it difficult to infer more specific topics which are only expressible by combining related concepts. Several studies have addressed this more so in utterance's extraction (Budak et al., 2014),incorporation of hierarchical relationships in knowledge bases such as Wikipedia (Kapanipathi et al., 2014b) as well as in extraction of persistent topics in tweet streams (Shin, Ryo & Park, 2014). It is also an issue in studies related to linkage of twitter posts and news articles for contextualization of Twitter activities (Abel, Gao, Houben & Tao, 2011d) and indirect examination of entities mentioned in Twitter posts for extraction as topics of interest (Michelson & Macskassy, 2010). This is addressed in Chapters 4 and 5.

3. **Inter-document correlations** with regard to the semantics is another fundamental area in improving the taste profiling process. Contextual understanding of the disseminated content is thus pertinent. Previous studies did not consider the context within which micro posts were disseminated but posts in the dataset were independently evaluated (Budak et al., 2014), (J. Wang et al., 2014), (Lu, Lam & Zhang, 2012), (Kapanipathi et al., 2014b), (Michelson & Macskassy, 2010), (Kapanipathi, Orlandi, Sheth & Passant, 2011). In time-based snapshots, this may just require a comparison with related posts at the time to discern the true meaning and extract user interests better. Research contributions in these aspects are made in Chapter 6.

4. **Short and long term taste profiling** is another area that has received attention from researchers (Abel, Gao, Houben & Tao, 2011a), (Budak et al., 2014). User interests and overall taste profiles, especially in short-text microblogs, change over time. To put this in perspective, users with a declared interest in, for example, a certain movie genre, are likely to watch and make their views known on other

genres over time. This evolution process from one area of interest to another is the best representation of user interest and subsequently, profile changes. Knowledge in evolution of tastes is contributed in Chapter 6.

The above literature denotes the overall research problems in modelling user-representative profiles in short and often noisy texts. Research gaps are identified and pointers to the approaches in this research contributing to knowledge on the same given. In addressing the research problem and gaps, contributions related to semantic understanding of short texts, extraction, representation and correlation of diverse user interests as well as modelling evolving taste profiles based on the interests are made.

1.3 Motivation of Research

Formulating user representative taste profiles with data that are short, noisy, sparse, streaming and with high dimensionality in their representation is an arduous task (W. X. Zhao, Jiang, Weng et al., 2011),(S.-H. Yang, Kolcz, Schlaikjer & Gupta, 2014a). At a granular level, the data have few descriptors that are specific to individuals and tailored for user-representative recommendations. Explicit user profiles information are sources of some metadata that could potentially be used in formulating taste profiles for the same users. On the contrary, such information is less likely to have statistical influence on the disseminated content. In some cases, such profile data are missing, or does not largely reflect the nature of actual tweets in terms of interests. Dynamism in the disseminated content too plays a part in rendering profile information insignificant. In short-text microblogs such as Twitter, the 80-20 rule, also known as the Pareto Rule, is real in the sense that 80% of content is generated by just 20% of the users on such platforms (Cha, Kwak, Rodriguez, Ahn & Moon, 2007). In addition, few users (influencers) on such streaming platforms have a large number of followers, whereas many users have few followers. This leads to the generation of data being limited at the

individual level, denoted as *sparse/thin* data. Concatenation of user-related data is one way to make thin data thicker, which is not easy based on the availability of minimal individual data (Zafarani & Liu, 2013; Korula & Lattanzi, 2014).

There is a need to develop algorithms that not only factor in semantic relations between terms and documents, but also fit noisy and shorter texts, which is a key focus of this research. Extraction of multiple interests in the short texts, as opposed to the single concept idea, is also a primary motivation for carrying out this research. This is equivalent to correlating concepts and documents. Additionally, formulation of more accurate and user representative taste profiles in short-text microblogs factoring in the extracted user interests is another motivation for carrying out this research. Variational time-stamp based snapshots of long and short-term interests in relation to the user/group social network structure informs this. Personalising user interests at each defined timestamp as a homogeneous interest graphs comprising of users, content and interests is also part of the profiling process.

To a large extent, this research is motivated by the ability to build taste profiling systems by extracting user interests from short and noisy texts. The deduction of interests is key in recommendation systems on such platforms, whether at individual or group levels. This process is fundamental in understanding and subsequently measuring audience based on their disseminated content. The output in this process is vital in third-party content and follow-back recommendations. The motivation in selecting short-text streaming platforms like Twitter in this research is with regard to their popularity in citizen journalism, content engagement in the user's follower-followee relationships, as well as the ability to integrate social theories with computational methods to determine human behavioral patterns and interests. This research is of interest in many fields, e.g., marketing, communication, e-democracy, and sports among others. This thesis covers different approaches in modelling user interests and formulation of user representative profiles. Thus, our approaches can be applied in diverse domains. Specifically, some of

our approaches contribute to fields such as sports betting and politics that are affected by the use of OSNs.

1.4 Research Objectives

This research focuses on the formulation of a framework for extraction of user representative taste profiles based on user-disseminated content, that is inherently noisy and short. As indicated in Sections 1.1 and, 1.3, accuracy in the taste profile formulation is fundamental. With the user representativeness in the profiling process being key, a solution is accomplished by achieving the four objectives listed below : -

1. **Objective 1**: To accurately extract notions/concepts from external sources to enhance learning and subsequent formulation of user interests in short and noisy texts.

   In the knowledge extraction process, the data deemed to be of high dimensionality is of interest. A model based on long-text external texts is made use of to guide the learning process. It is semi-supervised as it incorporates domain-specific metadata in the modelling process. The metamodel comprises of a set of topic label vectors derived from long texts to guide the learning process in shorter texts. This is the foundation of the research in modelling short texts for extraction of more interpretable interests as topics.

2. **Objective 2**: To formulate user interest levels in certain content based on their disseminated content.

   User interests profile is defined as a data structure representative of a group's/individual's degree of interest in a set of topics. Normally, the topics are in form of concepts or words (Piao & Breslin, 2018). In addressing this objective, the following steps were followed : - (a) Users' Degree of Interest (DoI) towards the identified

topic/concept based on the overall users' affinity towards that topic was computed. (b) An affirmation of the DoI was computed by correlating the DoIs to the user's friendship network based on the *theory of homophily* (Halberstam & Knight, 2016). A neural-network approach was utilised in the modelling process and computation of the interest levels in the topic.

3. **Objective 3**: To semantically detect user interest levels in varied content of interest/concepts. Objective 2 above centred around identification of users/group interest levels in one concept. It is applicable for example in the computation of a user's interest in a specific topic such as sports betting. An output of such a process is significant in identification of users for follow-back recommendations such as in a cold-start scenario. Objective 3 builds up on the same but in a multi-interest perspective. A combination of a neural-network and soft clustering (Peters, Crespo, Lingras & Weber, 2013) was integral in discerning the interest levels in different topics/concepts in the dataset.

4. **Objective 4**: To design, implement and evaluate a version of the framework that supports variational timestamps.

   Extraction of interests and subsequent building of taste profiles over variational time frames simulates changes in interests over time in the perspective of information gain and decay. To fulfil this research objective, a time-variant distributional model was created. The model is a fusion of words and topics/concepts over a period of time. A modified Dynamic Embedded Topic Model (D-ETM) (Dieng, Ruiz & Blei, 2019a) was created to learn short-text based embeddings segmented over time. This was for a class of users with a known seed topic. The approach makes use of Latent Dirichlet Allocation (LDA) (Blei et al., 2003a) to generate summaries of K preset topics through a discrete probability distribution over words. Topics are then distributed over neural-network based word embeddings

for better generalisation of topics. This process allows for a smooth variation of topics over long periods and the dataset. The model output was evaluated over actual news items at the specific times and a correlation matrix mapped as verification of an accurate representation of interests as true user taste profiles.

1.5 Research Contribution

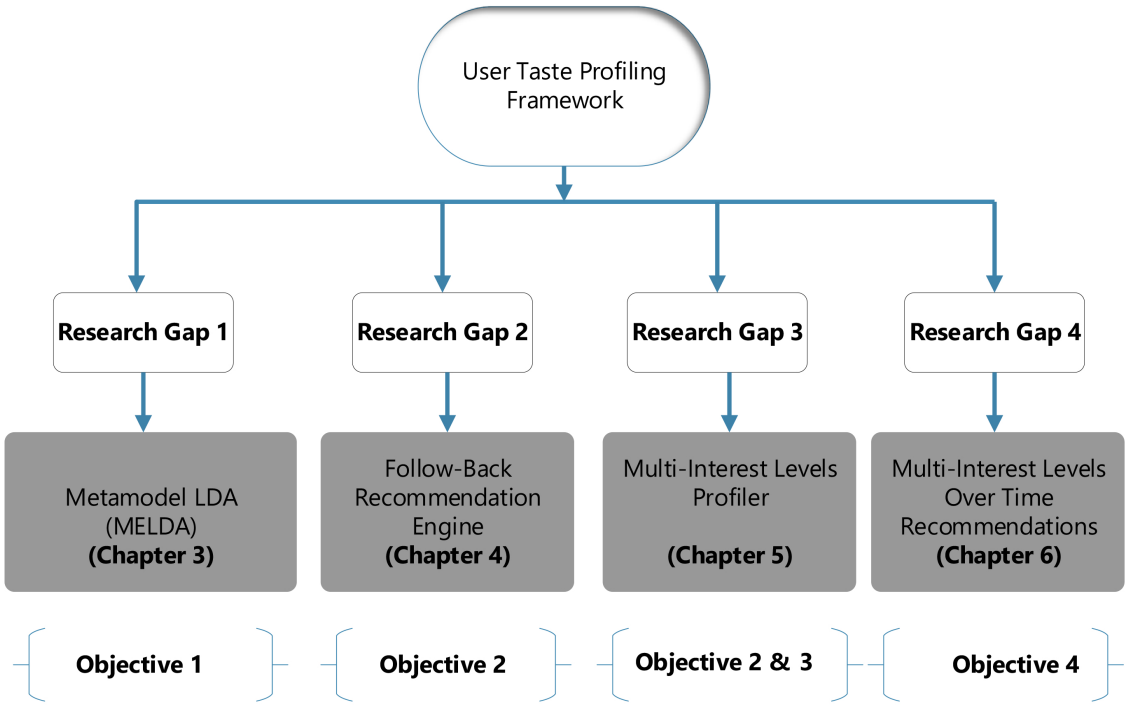


Figure 1.1: Overview of research solution

An overview of the complete research solution is presented in Figure 1.1. The framework is made up of four corresponding products that cover the research contributions: a Metamodel short-text knowledge extraction model (MELDA), a follow-back recommendation engine, a multi-interest profiler and a related multi-interest levels over time recommendation engine. The goal in these products is semantic knowledge extraction in short texts. These inter-related components contribute significantly to the the development of user-representative taste profiles.

In fulfilling the research objectives in Section 1.4, specific contributions were made in the quest to model user representative taste profiles on short-text microblogs. They are as below :-

- A semi-supervised approach called Metamodel-Enabled Latent Dirichlet Allocation (MELDA) that incorporates domain-specific metadata is implemented to guide the modelling process in short texts (Chapter 3).
- A modelling process that allows for the quantification of user interests in a certain concept/topic is implemented (Chapter 4). The interest measure is a factor in determining follow-back recommendations for certain users or used in content recommendations.
- A multi-interest user profiling process that allows for the quantification of user interests across several concepts /topics is implemented (Chapter 5). For example, based on the disseminated content across users, one may depict interests to a certain level per concept. A user's quantifiable interest levels may be 20% in *politics*, 40% in *sports*, 20% in *daily news* and 20% in *science*.
- In simulating information gain and decay, a common phenomenon in short-text microblogs, a combination of dynamic concepts and embeddings over time is done. This is meant to capture divergence of interests at specific times over the dataset, representing the changes in user interests over time (Chapter 6).

This work has produced the following publications:

1. **Wandabwa, Herman.**, Naeem, Muhammad. A., Pears, Russel., & Mirza, Farhaan. "A Metamodel Enabled Approach for Discovery of Coherent Topics in Short Text Microblogs," (2018), IEEE Access, 6, 65582-65593 (Chapter 3).

2. **Wandabwa, Herman.**, Naeem, Muhammad. A., Mirza, Farhaan., & Pears, Russel. "Follow-back Recommendations for Sports Bettors: A Twitter-based Approach," (2020, January), In Proceedings of the 53rd Hawaii International Conference on System Sciences (Chapter 4).

3. **Wandabwa, Herman.**, Naeem, Muhammad. A., Pears, Russel., & Mirza, Farhaan. "Topical Affinity in Short Text Microblogs," (2020), Information Systems Journal (Chapter 4).

4. **Wandabwa, Herman.**, Naeem, Muhammad. A., Pears, Russel., & Mirza., Farhaan, Andy., Nguyen. "Multi-Interest User Profiling in Short Text Microblogs," (2020). Submitted to the 15th International Conference on Design Science Research in Information Systems and Technology (DESRIST). Published in Volume 12388 of the Lecture Notes in Computer Science series, Extending the Boundaries of Design Science Theory and Practice (Chapter 5).

5. **Wandabwa, Herman.**, Naeem, Muhammad. A., Pears, Russel., & Mirza., Farhaan. "Multi-interest Semantic Changes overtime in Short Text Microblogs,"(2020), Submitted to the Knowledge-Based Systems Journal (Chapter 6).

6. **Wandabwa, Herman.**, Naeem, Muhammad. A., Pears, Russel., & Mirza., Farhaan. "User Representative Profiling in Short-Text Microblogs - Concepts, Challenges, and Opportunities (2021), Submitted to ACM Computing Surveys (Chapter 2).

1.6 Thesis Structure

The structure of rest of the thesis is presented as follows: -

Literature related to user interests/preferences extraction in Online Social Networks (OSNs) is presented in Chapter 2. Concepts that led to this research more so in areas of personalisation, content recommendations and *user profiles modelling* are introduced. Since this research is based on short and noisy texts, modelling approaches in this domain and in relation to the aforementioned research areas are also detailed.

Chapter 3 introduces short-text modelling including the motivation behind the choice of short-text data. In an attempt to improve on the state-of-the art (SOTA) in this domain, domain-specific conventional data were incorporated in order to generate more interpretable concepts/topics for the short texts. Therefore, the modelling process approach is semi-supervised. Since the modelling goal is to improve the accuracy in the generated topics/concepts, topic coherency as the output metric is higher than other comparative semi-supervised and unsupervised SOTA approaches. The goal of this approach is to extract semantically relevant concepts in a dataset with noisy and short texts for further modelling.

In Chapter 4, the choice of representing text as vectors at character level is made. This is because neural network based embeddings that model words atomically do not work well for noisy texts. Therefore, FastText-based vector representations are used in modelling textual data in the experiments. Specifically, user interest levels are modelled with respect to a specific concept/topic by (i) computing the Degree of Interest (DoI) for users towards the concept that is identified as a cluster centroid. User affinity towards identified concepts is the end result. (ii) An affirmation of the DoI is then computed by correlating with the friendship network's DoIs based on the *homophily theory* (McPherson, Smith-Lovin & Cook, 2001).

Chapter 5 complements the work presented in Chapter 4 as multi-interests representative of divergent user interests are modelled in this chapter. The divergence is in terms of interest levels exhibited towards different topics/concepts of interest. To model this, soft clusters are computed over embeddings by (i) converting all tweets to vector representations; (ii) extracting cluster centroids representative of the concepts/topics of interest; (iii) computing responsibility matrix depicting user interest levels in the concepts; (iv) aggregating intra-user interest levels to define the user's multi-topic affinities.

In Chapter 6, a framework that simulates a typical short-text dissemination platform is presented. This is in the aspects of information *decay* and *gain*. This is compounded by the rate at which certain topics gain and lose attention across timestamps. In presenting this in terms of user profiles, time is factored in the embedding space. A variant of the Dynamic Embedded Topic Model (D-ETM) (Dieng et al., 2019a) is utilised to learn short-text based embeddings segmented over time for a class of users identified in Chapter 4.

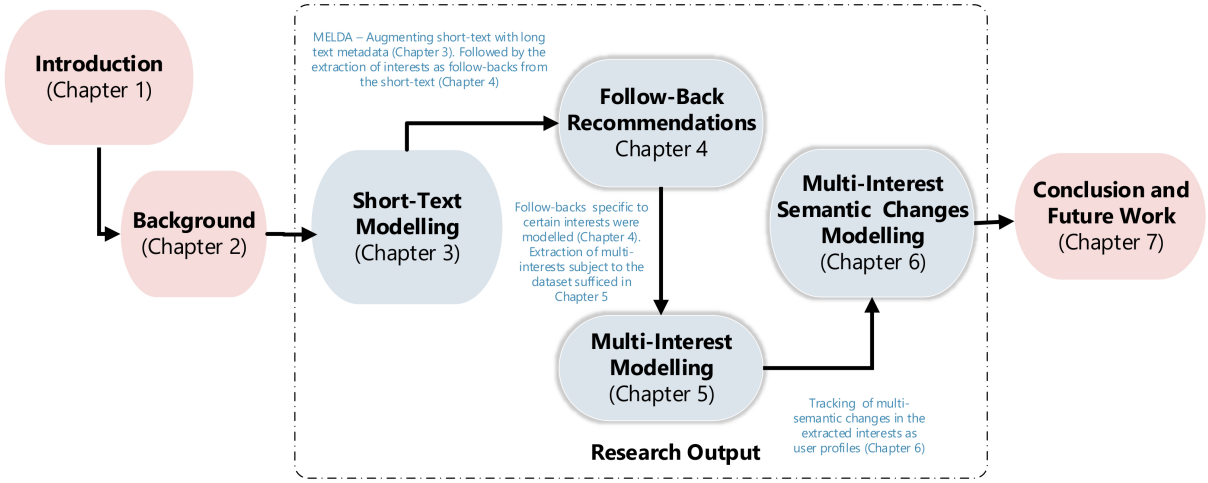


Figure 1.2: Conceptual Thesis Roadmap

(Chapter 7), summarises the research and details how the objectives of this thesis were achieved. Furthermore, future research directions in taste profiling and final

remarks are also outlined. Figure 1.2 illustrates the conceptual roadmap of the thesis.

Chapter 2

Related Work

2.1 Introduction

This chapter provides a summary of the related literature that forms the basis of this research. The discussion starts with definitions of a few concepts related to modelling of short and long texts, as they are the foundations of this research. Thereafter, the challenge of user interests extraction, modelling and taste profiling in short texts is explored. Dynamism in the profiling process is also explored. This is important in the replication of this research in an actual short-text streaming environment. Data throughput in short-text microblogs is enormous, meaning that some topics gain attention and decay at a very fast rate according to the prevailing discussion at the time. Fundamentally, this research addresses the problem of extracting user-representative profiles in text that is not only noisy, but whose topical gain and decay are quite dynamic. Therefore, approaches related to extraction of knowledge in short texts and especially over periods in building taste profiles are vital in this research.

In order to address issues related to noisy and streaming texts as well as conventional long texts in modelling of user representative profiles, researchers have applied various approaches. To highlight what other researchers have done in these areas, the research

facets listed below are of interest:

- Online Social Networks (OSNs)
- Modelling, knowledge extraction and recommendations in texts.
- Semantic User interests Mining.
- Taste profiling in microblogs.

A representation of areas that have been covered in this chapter is in Figure 2.1.

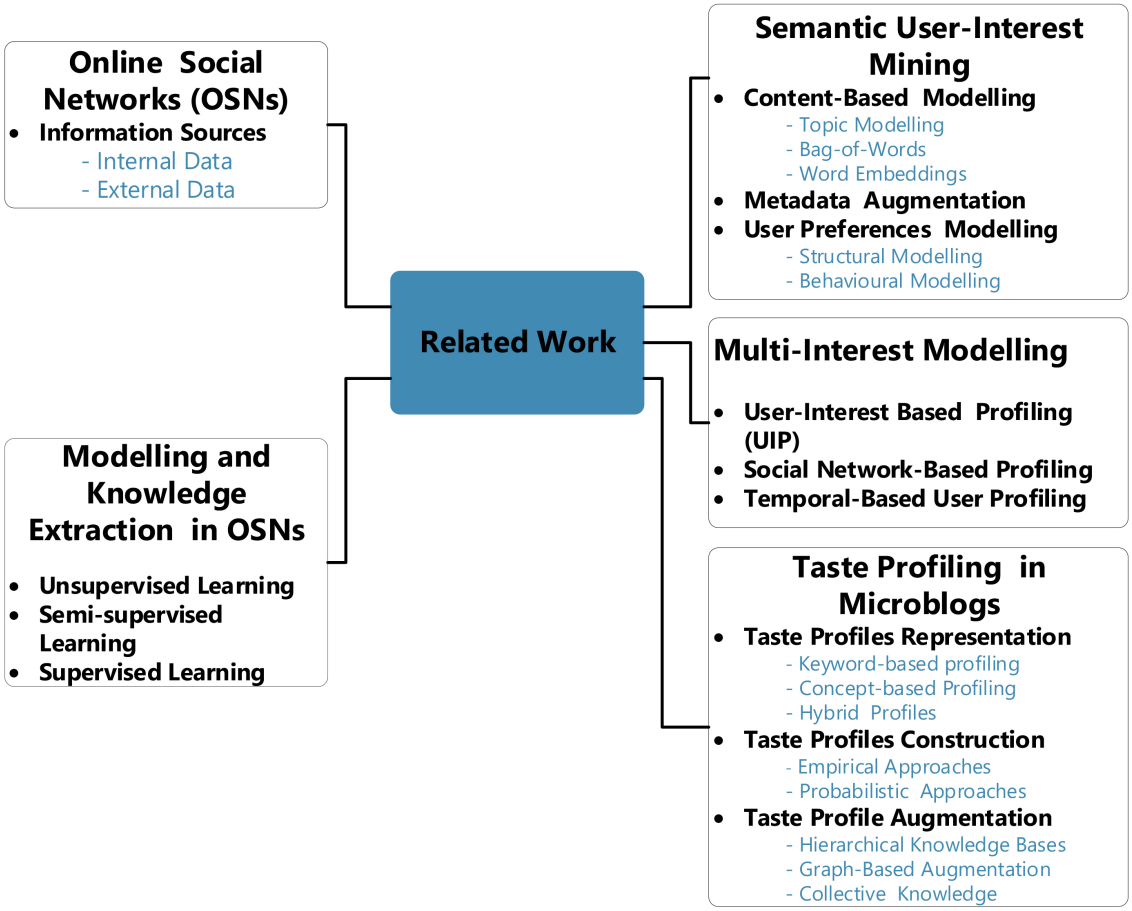


Figure 2.1: Related Work Overview

2.2 Online Social Networks (OSNs)

Online Social Networks (OSNs) as defined by Boyd et al., (Boyd & Ellison, 2007) are web-based applications and related services that (i) explicitly allow for the construction of a public or semi-public profile within their system; (ii) list other users with existing or potential relations with the original user; (iii) allow the individuals to navigate their list of connections. The connection types vary per OSN. OSNs have revolutionised the way humans think and interact as they support social cognition in humans (Chi, 2008). Apart from individuals on such platforms looking to connect with new people, they also help users to maintain relationships with current contacts as part of their network (Boyd & Ellison, 2007). This is notable in created groups on such platforms.

There exist different types of microblogs with a definitive difference in the nature of user connections on the platforms. Some platforms such as Facebook[1] expect bidirectional relations for users to communicate. On the other hand, user relationships on Twitter[2] can either be uni-directional or bidirectional. The common feature among them is in the interactions among users on the platforms (Kraut & Resnick, 2012). Research in OSNs has centred around mining and modelling User Generated Content (UGCs) on the platforms as well as in the exploration of social connections for collective intelligence or extraction of knowledge for decision making, etc.

2.2.1 Information Sources in OSNs

Information sources for collecting data in OSNs for inferring user interests are varied. They can either be (i) *internal,* meaning they can be collected from the OSN platforms, e.g., tweets, or (ii) *external,* where the data are collected from external sources, e.g., external databases or ontologies. External data can be used to augment modelling of

[1] `https://www.facebook.com/`
[2] `https://www.twitter.com`

internal data.

Internal Data

Microblog users have the option of giving more descriptive information in their profiles. This is based on the extent to which they explicitly express themselves. A user's biography (bio) where one's listed interests and hobbies are informative in user interest modelling (Piao & Breslin, 2017a). On the contrary, a bulk of the users provided profile information that may not be semantically representative of the user, as it may not be overly accurate, because of evolution of interests over time. The most representative user interests profile relates to the content the same users disseminate, like or retweet, e.g., on Twitter. The posts are typically made up of terms that are semantically relevant to a certain concept. For example, if a user often mentions "Trump" in his/her posts, there is a high chance that "Trump" and "politics" may be of interest to the user. Extraction of semantic concepts in form of features such as named entities in posts is an area that has been explored (Z. Zhao, Cheng, Hong & Chi, 2015).

Embedded links (URLs) and tags are significant in microblog content (Peñas, Del Hoyo, Vea-Murguía, González & Mayo, 2013; Piao & Breslin, 2016b). They are vital in the inference of user interests, e.g., hashtags on Twitter. These are terms representative of a certain topic/concept that is of importance at the time. What differentiates them from other words is that they are usually preceded by the hash symbol "#". The same approach has been replicated in photo and video-centric microblogs like Instagram, Flickr, and Tumblr, where images or videos are pinned to messages. Such multimedia content has proved to be a good source for mining user interests (Joshi, Cooper, Chen & Chen, 2015; Grbovic, Radosavljevic, Djuric, Bhamidipati & Nagarajan, 2016).

Lastly, internal data in the form of *List membership of followees* on Twitter has been studied as a user interests source (Piao & Breslin, 2017b). *Tweeters* can freely create

lists and add users with interest in the list's topical content. For example, users with an interest in data science can be added to a group called "Data Science". Therefore, followers of members in this list can be inferred to also have an interest in the data science related content.

External Data

Short-text microblogs are short and noisy, and thus present knowledge extraction challenges compared to conventional texts. In addition to the works mentioned in Section 2.4.1, external knowledge bases and independent news articles have been leveraged to augment the knowledge extraction in tweets (Bontcheva & Rout, 2014; Abel et al., 2011d; Piao & Breslin, 2016b). Here, tweets were linked to news articles for semantic enrichment. Findings in the research by Kwak et al., (Kwak, Lee, Park & Moon, 2010) indicated that 85% of tweets were news related.

External knowledge bases such as WordNet and Wikipedia have been incorporated in semantic enrichment of short texts. The knowledge bases provide explicit semantic description of concepts/categories and their relationships. This is instrumental in semantically contextualising short texts that are otherwise short and noisy (Kapanipathi et al., 2014b; Michelson & Macskassy, 2010).

To develop and evaluate methodologies related to short and noisy texts in this thesis, Twitter datasets were made use of. Thus, a description of Twitter as a short-text microblog follows.

2.2.2 Twitter

Introduction

Twitter, a microblogging OSN that proliferates digital data due its streaming nature is the short-text OSN of choice in this research. Therefore, a brief description of the same

is of essence. Twitter is instrumental in sharing near to real-time data in the form of text, videos, hyperlinks and images. Activities on Twitter are depicted in the form of tweets, retweets, replies, likes and shares. The user connection structure on the platform is based on *follower-followee* relationships in a unidirectional or bidirectional manner. Initially, tweets were made up of 140 characters, but this limitation has since been adjusted to 280 characters. A tweet's metadata consists of the text, pictures, date of dissemination, geo-coordinates, URLs, hashtags, and mentions, among others.

Users on Twitter (tweeters) follow their friends, influencers, and other pages of interest. With such filters and in addition to the implicit ones such as geographical location, tweeters are able to access a filtered timeline of their networks' tweets. To promote such content, tweeters are able to re-share the disseminated tweets as "retweets", as well as "comment" on and/or "like" the original tweet, etc. This content propagation process is instrumental in citizen journalism. Statistically, the number of disseminated tweets averages about 500 billion per year, which roughly translates to approximately 6,000 tweets per second[3].

Twitter Lingo

Twitter just like other short-text dissemination platform, has a language of its own. For example, some symbols depict certain actions e.g. the @ sign, that symbolizes a Twitter handle or username on the platform. Below are major expressions/definitions specific to Twitter : -

- ***Tweets*** - These are the messages that are disseminated on the Twitter platform. They are limited to 280 characters thus the short-text classification. They can comprise of text, videos, photos or links, and are are shareable. A sample tweet is shown in Figure 2.2.

[3] `http://www.internetlivestats.com/twitter-statistics/`

Figure 2.2: Sample tweet

- ***Retweets*** - Ideally, a retweet just expresses emphasis when a tweeter re-shares a tweet directly or quotes the tweet in his/her tweet. This is achieved by clicking on the retweet button at the end of the disseminated tweet.

- ***Hashtags*** - Hashtags are keywords or phrases that are representative of a topic and are preceded by the # symbol on Twitter. Hashtags connect conversations and make it easier to find content. This way, communities are able to converse around a topic easily, by referring to the hashtag.

- ***Mentions*** - The @ symbol extracts usernames that are unique identifiers of tweeters. The keyword immediately after the @ symbol is a user's unique identifier on the platform. For example, searching @POTUS[4] directs one to the US presidents page on the platform. The rest of the lingo specific to Twitter is found here[5].

[4] https://twitter.com/POTUS
[5] https://help.twitter.com/en/glossary

Collection of Tweets via The Twitter API

Twitter has an API endpoint that can be used to programmatically retrieve, analyze data as well as engage in conversations on the platform. The API provisions access to resources such as access to a variety of resources including tweets themselves, users, messages, lists, trends, places and the shared media. A request for the API is typically done via an external application after creating a developer account. Once the account is approved, the user is required to create a project and app, that will enable for the generation of the below connection credentials : -

- ***API Key*** - This is the username that allows one to make a request on behalf of the app.
- ***API Key Secret*** - This is the password that allows one to make a request on behalf of the app.
- ***Access Token*** - This token is representative of the Twitter account that owns the app, and allows one to make a request on behalf of that Twitter account.
- ***Access Token Secret*** - This credential also represents the Twitter account that owns the app, and allows one to make a request on behalf of that Twitter account.
- ***Bearer Token*** - This represents the app and enables one to authenticate requests that require OAuth 2.0 Bearer Token [6].

The API Key, API Key Secret, Access Token, and Access Token Secrets are used to make requests that require OAuth 1.0a User Context authentication [7]. More details especially on the access setup are found here [8].

[6] https://developer.twitter.com/en/docs/authentication/oauth-2-0authentication

[7] https://developer.twitter.com/en/docs/authentication/oauth-1-0a

[8] https://developer.twitter.com/en/docs/twitter-api/getting-started/getting-access-to-the-twitter-api

Profile Generation and Augmentation

A tweeter's profile definitively identifies their interests. A profile on Twitter entails getting access to identifiable information that extrinsically describes the user. Normally, users will have a profile picture and an introduction in the form of a bio as well as a description of their interests. Other details include the user's name, location, website, phone number and birth date. Profiles on short text platforms should be dynamic to match the changes in interests. For example, Twitter encourages users to continually change their profile information to match what truly reflects their status at the moment.

Intrinsically, a user's disseminated content i.e., the tweets are a precursor to the generated profile. They are deemed to be a true representation of the user's interests at that time. Therefore, an amalgamation of intrinsic and extrinsic user information on such platforms is bound to present user profiles that are optimally representative of the user, more so in their declared and undeclared interests.

Social Aspects of tweets - Followership

Users on short-text platforms are bound to be interested in content/topics as well with other users that are relevant to them. When such users or topics are found, one is likely to *follow* them. This forms a unidirectional followership as it is one-sided. If the followed user decides to follow the followee, then the relationship will be bi-directional. Therefore, users in such a relationship will be able to view each others disseminated content. Their followership provides a link that can be used to recommend relevant content/users to their friendship network. To follow a user, one simply needs to click on the follow button next to a user's profile. Clicking of the same button results in unfollowing the same user.

Followership in topics is unidirectional. Normally, this will happen when tweeters want to see more content about topics relevant to people they are interested in, but do

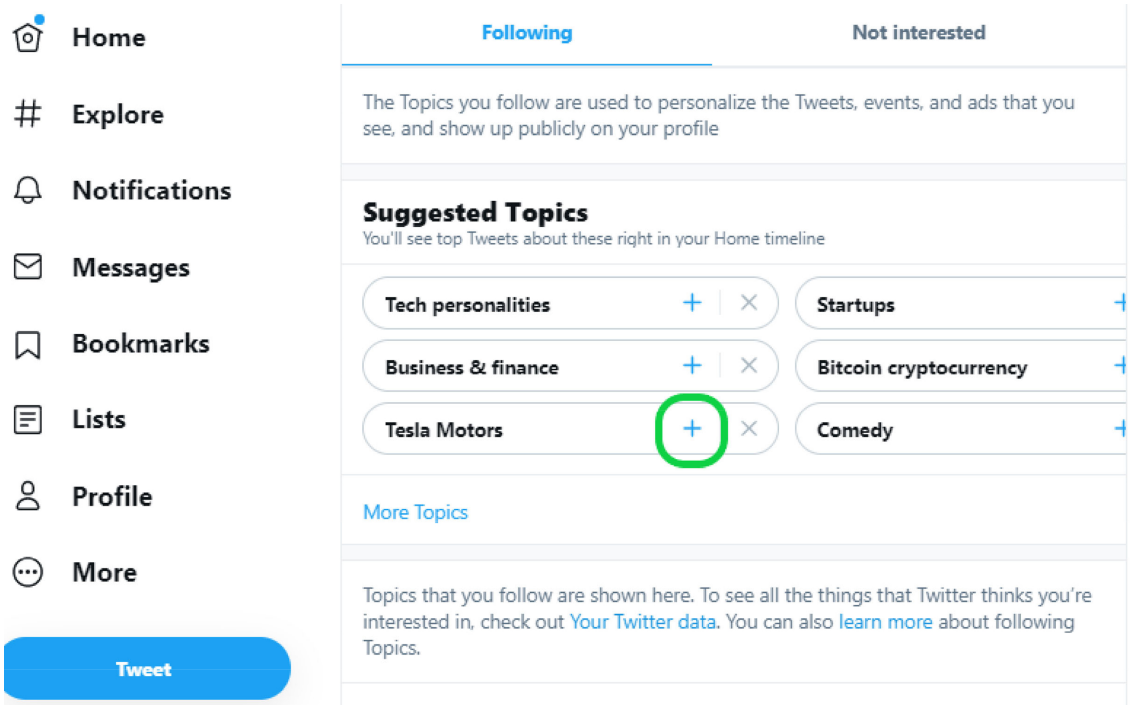


Figure 2.3: Suggested topics of interest for a user.

not want to necessarily follow the users. A list of suggested topics as shown in the Figure 2.3 will be displayed when the topics tab is clicked. Following a topic entails clicking the plus (+) sign next to the topic. This way, users with the help of Twitter, are able to curate the content they consume based on their optimal tastes. In case one is not interested in the suggested topics, then they can click on the X as in Figure 2.4. This triggers the platforms recommendation engine not suggest the topic in future.

Overall, the ***theory of homophily*** is actualised in social networks, more so on Twitter with the above mentioned relationship entities in terms of users and, topics followership (Halberstam & Knight, 2016). Ideally, this is the same principle of "birds of the same feather", where users with either unidirectional or bidirectional relationship on the platform are assumed to have commonality in their interests. Therefore, the assumption is that users with interest in *rugby* or *soccer* are likely to be friends with users whose interest is in the same domain.

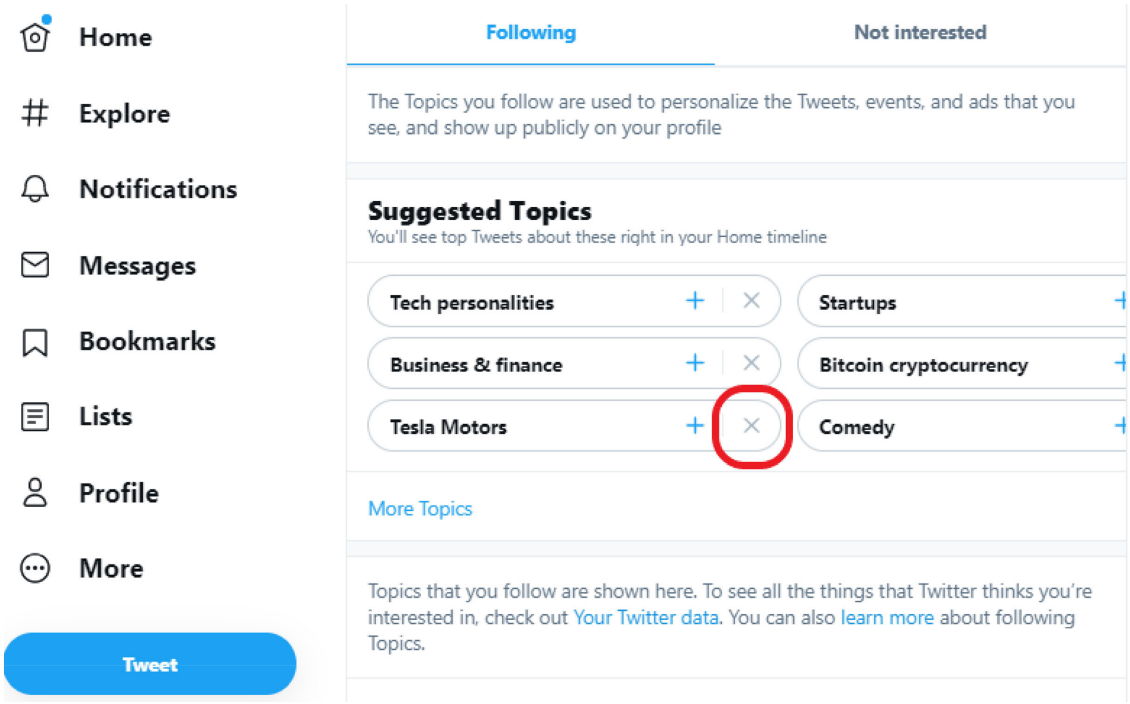


Figure 2.4: Suggested topics of interest for a user.

Tweets Dissemination, Information Gain and Decay

Short-text on microblogs such as Twitter, gain and lose attention as engagement levels vary over time. Below are a selection of few features affecting dissemination patterns on the platform: -

1. **Engagement** - Throughput on Twitter is high. This is based on the engagement patterns as depicted in various approaches in Section 2.2.2. Therefore, engagement on the platform is important. For better dissemination on the platform, the disseminators are supposed to actively engage the audience by e.g. replying to tweets and messages. This is paramount for influencers to maximize their network reach. In addition, retweets and sharing of content from other users, promotes engagement in addition to curating a relevant user's profile. Additionally, Twitter provisions for users to go live and share experiences as they happen. New and engaging content coupled with mentions of users in the network, more so

Figure 2.5: Sample screenshot of trends in Nairobi, Kenya on 20th April 2021 at 1:34 PM(GMT+3)

influencers, is bound to invoke interest from other users.

2. **Trends** - Trends or trending topics are simply topics that have gained attention in a short timeframe. This can depend on the user's location, meaning the user will have access to trending topics in their area of interest. For example, one can be in Auckland, New Zealand, but set access to trends in Nairobi. This results in one accessing topics of engagement in Nairobi. The reverse is also possible. A sample screenshot of the trends is in Figure 2.5.

3. **Lists** - Lists on Twitter allow users to organize tweets better on their timelines. Users can join lists or create some themselves and, have people join. Lists can be created based on the topics of interest. A list timeline displays tweets specific to the accounts in the list, that can also be pinned on the timeline to maximize reach. Lists and all other facets of engagement on the platform are meant to augment intrinsic profile details. A sample lists dashboard is shown in Figure 2.6.

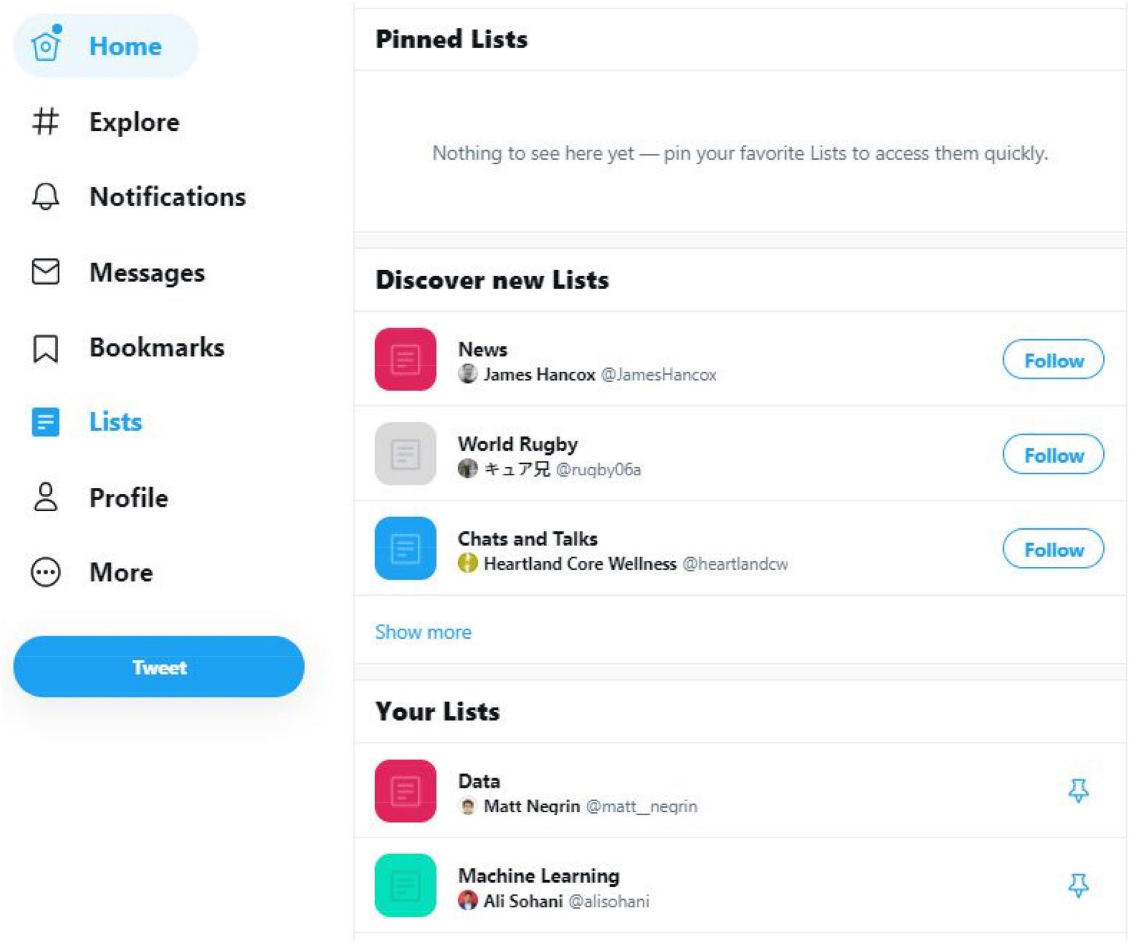


Figure 2.6: Twitter dashboard with lists of interest.

The above features are largely unique to short-text microblogs and specifically to Twitter. The platform's uniqueness is predominantly in the throughput and constant evolving of the disseminated information as well as the nature of the generated data. Character limitations call for creativity in the tweets dissemination patterns, making it possible to learn social exploits for tweeters. This is compounded by the connection patterns as relationships with users and topics on the platform.

2.3 Modelling, Knowledge Extraction and Recommendations in Texts

As much as unstructured textual data can be easily processed by humans, it is significantly difficult for machines to process the same. *Text mining* is thus fundamental in this research, more so in short texts. Text mining is basically the process of extracting information from structured sources, e.g., Relational Database Management System (RDBMs), and semi-structured or unstructured sources (Witten & Frank, 2002). Techniques and algorithms related to clustering and topic modelling are grouped as unstructured algorithms in textual data (Maimon & Rokach, 2005).

Learning textual patterns is primarily dependent on the type of data in question. Longer texts, e.g., products and services reviews, text in email messages, etc., form the bulk of online textual data. The notion is that conventional texts have better word co-occurrence patterns, as terms are often repeated, and thus statistically significant. In most cases, *supervised* and *unsupervised* methods are used to extract knowledge in such sets. The same methods can be applied on shorter texts, but have limitations due to the challenges related to bias, data insufficiency, data noisiness, big data paradox, and a lack of proper evaluation techniques in microblogs. The knowledge extraction processes in text data is classified as follows, based on the nature of the data:

2.3.1 Unsupervised Learning

Unsupervised methods are knowledge discovery techniques, mostly applied to unlabelled data. In situations where, for example, the data are too huge or streaming in nature that one cannot manually annotate, then such an approach will suffice. *Clustering* and *topic modelling* are two commonly used approaches, especially in textual data context. In clustering, the collection of documents is segmented into partitions where

semantically close documents are grouped together as compared to distant ones. On the other hand, topic modelling is probabilistic, whereby every document has a probability distribution over the other document clusters in the collection (Aggarwal & Zhai, 2012). Each topic is represented as a probability distribution over words and each document as a distribution over topics. A topic is equivalent to a cluster, and membership of a document to a topic is probabilistic (Steyvers & Griffiths, 2007).

2.3.2 Semi-Supervised Learning

In semi-supervised learning, the usage of both *labelled* and *unlabelled* data is pertinent owing to the practicality involved in labelling data, particularly with big data. For example, there is an abundance of unlabelled data in short-text microblogs, e.g., tweets. This is because of the impracticality in labelling for example, 2 billion tweets. Semi-supervision in such a scenario involves labelling part of the data to formulate a link between the unlabelled data distribution and, for example, the classification function. Methods related to generative models; semi-supervised support vector machines, utilise this approach (Chapelle, Scholkopf & Zien, 2009).

2.3.3 Supervised Learning

On the other hand, mapping a set of input variables A to output variables *B* characterises supervised learning (Cunningham, Cord & Delany, 2008). This mapping is applied in the prediction process on unseen data. Exclusive availability of annotated data differentiates this approach from those in Sections 2.3.1 and 2.3.2.

2.4 Semantic User Interest Mining

User interests from the perspective of this research, are specific aspects of importance in the identification of user(s) activities on short-text microblogs. Extraction of user interests from user's social data is important in many applications; from third-party recommender systems to homophily analysis. Usually, third-party content and service providers can significantly reduce audience reach costs by offering the most relevant products (e.g., ads and related third-party content) to their customers. However, accuracy in the identification of user interests is a challenge (Zarrinkalam, Fani & Bagheri, 2019).

In microposts such as tweets, users can express their feelings and views in real time. This presents such microblogs as sources of explicit and implicit information for inference of users' interests (Budak et al., 2014), (Shin et al., 2014). Filtering of twitter streams (Bontcheva & Rout, 2014), (Pennacchiotti, Silvestri, Vahabi & Venturini, 2012), identification of social communities (Palsetia, Patwary, Agrawal & Choudhary, 2014), news recommendations (Abel et al., 2011a), among others are also potential application areas. There are several approaches used in extraction of user interests from textual data:

2.4.1 Content-Based

Content-based approaches in user-interests extraction encompass the analysis of the textual content disseminated by microblog users. Among other techniques, the following are dominant as mostly the textual aspect of what is disseminated is of interest:

Topic Modelling

As introduced in Section 2.3.1, topic modelling is probabilistic in nature. This means that topics are curated by extracting groups of co-occurring terms and each document is viewed as a mixture of various topics (Steyvers & Griffiths, 2007). Latent Dirichlet

Allocation (LDA) is such an approach used in the detection of user interests in textual content (Ramage, Dumais & Liebling, 2010). Weng et al., (Weng, Lim, Jiang & He, 2010b) for example aggregated tweets as one large document and modelled the document via LDA as a longer document in dealing with the vocabulary sparsity issue. Labelled LDA which is a supervised version of LDA was used in the detection of user interests on user-specific tweets (Ottoni et al., 2014).

Bag-of-Words (BoW)

In the Bag-of-Words (BoW) approach, a set of terms representing user interests are extracted from texts (J. Chen, Nairn, Nelson, Bernstein & Chi, 2010). Yang et al.,(L. Yang, Sun, Zhang & Mei, 2012) modelled user interests by weighting term vectors and measured user similarity via cosine distance. Yongwook et al., (Shin et al., 2014) considered detection of topics as long-term steady interests to a user from a text stream. The authors considered streams dynamism and social characteristics via a graph-based topic extraction model.

Word Embeddings

Embedding words as vector representations has been used by researchers to extract contextual knowledge in textual data. In the approaches, the assumption is that semantically close words are likely to share similarities in their representations. The main idea in embeddings is in the generalisation power, whereby some features provide related clues (Goldberg, 2017).

The above approaches work well when term co-occurrence is high and sufficient samples are provided. However, they fall short when it comes to factoring in underlying semantics and the relationship between terms. In the absence of sufficient word co-occurrences reminiscent of short texts such as tweets, the approaches may not perform

well (Sriram, Fuhry, Demir, Ferhatosmanoglu & Demirbas, 2010), (Kapanipathi et al., 2014b) and (Michelson & Macskassy, 2010).

To address some of the shortcomings in the above approaches, *Bag-of-Concepts* (BOC) was introduced, whereby candidate concepts were extracted from external sources, e.g., Wikipedia, Freebase, and Yago. Such knowledge bases represent concepts and interrelations between them, thus to some extent provide an understanding of the underlying semantics. A *Twopics* approach was also proposed (Michelson & Macskassy, 2010). In the approach, a set of Wikipedia entities from a user's timeline are extracted. User interests are then identified by traversing their Wikipedia categories. DBpedia[9] concepts were also used to annotate tweets in modelling user interests. The annotations were then used as filters for the tweets (Kapanipathi et al., 2011). The authors introduced weighted primitive interests (bag-of-concepts) derived from entities in tweets and implicit interests that were extracted by mapping the primitive interests to a Wikipedia category hierarchy. Apart from Wikipedia, related news articles were used to enrich tweets in order to extract user interests from the enriched set (Abel et al., 2011d). However, the temporal dynamism in microblogs makes it difficult to pin posts to concepts via such external datasets.

2.4.2 Metadata Augmentation

Conventional long texts, e.g., products reviews, as well as short and noisy texts form the bulk of the online data. Comprehension of such texts is thus an active research area. Topic modelling algorithms such as PLSA (Hofmann, 1999) and LDA (Blei et al., 2003a) have demonstrated great success in the discovery of latent topics in large text corpora, especially in long text documents. However, the same algorithms do not perform well when subjected to short and noisy datasets (S.-H. Yang, Kolcz, Schlaikjer

[9] https://wiki.dbpedia.org/

& Gupta, 2014b; W. X. Zhao, Jiang, Weng et al., 2011). Despite this drawback, LDA remains one of the most popular algorithms of choice in uncovering latent semantic structures in such texts (S.-H. Yang et al., 2014b; W. X. Zhao, Jiang, Weng et al., 2011).

Regarding *modelling of topics* in short texts, Wayne Xin et al., (W. X. Zhao, Jiang, He et al., 2011) proposed a topic key phrases extraction methodology for tweets. In their approach, three processes were applied: keyword ranking, generation and ranking of key phrases. In ranking keywords, a modification of the Topical PageRank method (Z. Liu et al., 2010) was made by introducing a score related to topic sensitivity. Principled probabilistic phrase ranking was then proposed to rank key phrases. User interests were also modelled via the retweeting patterns using this key phrase ranking methodology. Experiments on a larger Twitter dataset depicted better performance in the extraction of topical key phrases in short texts. On the other hand, Zhao et al., (W. X. Zhao, Jiang, Weng et al., 2011) modelled topics from tweets by assuming that each tweet belonged to one topic, which is not overly true. A tweet such as *"Illiteracy is a product of a failed government"* depicts more than one topic, i.e., *education* and *governance*. Aligning the contextual representation of the tweet to one topic (single idea concept) was therefore not a viable hypothesis. Weng et al., (Weng et al., 2010a), proposed finding topic-sensitive influential tweeters by combining all tweets from one individual into one document. The goal was to understand each user's interests and LDA was applied on the single aggregated document. This approach, however, did not reduce noise, in addition to making word co-occurrence more difficult to interpret.

Incorporation of external knowledge in augmentation of short texts is another thought process that some researchers adopted in learning texts with a sparse vocabulary. Andrzejewski et al., (Andrzejewski et al., 2009) incorporated domain knowledge in the conventional LDA modelling process via Dirichlet Forest priors (DF-LDA). Their thought process was that knowledge could be expressed with two primitives on word pairs, i.e., Must-Links (i.e., words that should co-occur) and Cannot-Links (i.e., words

that should not co-occur). This was meant to restrain topic distribution. For example, words like "study, computing" or "Maori, culture" are guaranteed to co-occur in the topic modelling process. This vital co-occurrence knowledge was encoded as Must-Links in the Dirichlet Forest prior to augment the LDA process for extraction of succinct topics.

Chen et al., (Z. Chen et al., 2013b) proposed Multi-Domain LDA where s-sets (semantic sets) denoted sets of words sharing the same semantic meaning in a domain, similar to must-links in (Andrzejewski et al., 2009). Prior knowledge from multiple past domains is explored as s-sets to produce more coherent topics. Domain knowledge extraction from Wordnet (Miller, 1995) was also applied to extract knowledge from texts by Chen et al., (Z. Chen et al., 2013c).

In conventional long texts, Ramage et al., (Ramage, Hall, Nallapati & Manning, 2009) improved the topic modelling process by introducing known labels to produce even better topic labels through multi-label supervision. However, finding such prior knowledge, especially in streaming noisy texts such as tweets, may not be possible, due to their large topical variance. Incorrect external knowledge, on the other hand, may lead to unreliable results due to contextual variances in the two sets. For example, topics with well-defined descriptions may have been incorrectly influenced by prior knowledge and not by patterns in the actual data.

To mitigate the above user-centric problems, some researchers have taken advantage of underlying knowledge in the data. In *social text analysis*, social relationships between entities are one path that has been explored to find prior knowledge (Mei, Cai, Zhang & Zhai, 2008). Wang et al., (X. Wang et al., 2011) represented hashtags as graph values and mined co-occurrence patterns between them. Furthermore, the literal meaning of hashtags was utilised as semi-supervision.

Implicit information mining in social networks is another area that has been explored for knowledge extraction (He, Wang & Jiang, 2015). Based on the underlying social network structure, reposting the behaviour of users could also be modelled. This is

useful in the prediction of user content dissemination patterns that are pertinent in scenarios related to events discovery (Piña-García et al., n.d.). Combination of re-tweeting networks and textual content is also an interesting dimension in social text analysis. Community detection in social networks is also another area that has been extensively researched on in augmenting the knowledge discovery process in short-text microblogs (Pathak, DeLong, Banerjee & Erickson, 2008).

2.4.3 User Preferences Modelling

User interests are major contributors to the content that microblog users are prone to consume or disseminate, in addition to their being integral in the design of recommender systems. Recommender systems were developed to match users to the resources they are most interested in. This could be in the form of content or other users in cases of follow-back frameworks such as Twitter. There are two main approaches in the formulation of user profiles in recommender systems, i.e., *structural*(Symeonidis, Nanopoulos & Manolopoulos, 2007),(Jawaheer, Weller & Kostkova, 2014) and *behavioural*(Yin, Cui, Chen, Hu & Zhou, 2015a) modelling.

Structural modelling approaches are primarily used to extract features from or to stereotype a group of users (Brusilovsky & Millán, 2007). In essence, the formulated user profile is structurally representative of the user(s)(Brusilovsky & Millán, 2007). In collaborative filtering for example, a user profile can be modelled as a feature vector representing specific aspects of the user as entity descriptors, e.g., geo-location in the metadata of a tweet. Similarity-based algorithms (H. Liu, Hu, Mian, Tian & Zhu, 2014) can then be applied to measure similarity among users (Cai et al., 2013). In the context of content-based recommender systems, a keyword-based vector represents the user profile. A similarity measure is then used to match the distance between the profile and other resources (Aggarwal, 2016).

Behavioural modelling on the other hand identifies observable parameters that reflect user behaviour. This can be to a certain level of certainty, and thus are probabilistic, e.g., Bayesian networks (Yin et al., 2015a). An extension of the Latent Dirichlet Allocation (LDA) probabilistic approach, called Forum-LDA, was formulated to extract useful interest topics in online forums by modelling the generative process of seed posts as well relevant and irrelevant responses to the seed posts using a Bernoulli distribution (C. Chen & Ren, 2017).

In the light of recommendations in short-text microblogs, Chen et al., proposed a tweet's recommendation system based on collaborative ranking to capture personal interests by integrating useful contextual information such as tweet topic level factors (K. Chen et al., 2012). Authors in (Goel & Kumar, 2018) proposed a methodology for User Interest Profile (UIP) design by enriching tags generated by the user with friendship information via vector representations. Profiling malicious users in short-text microblog users is also an area of research interest. Authors in (Sahoo & Gupta, 2019) proposed a hybrid approach leveraging classifications and Petri net structure to profile malicious users on Twitter. On the other hand, researchers in (J. Chen et al., 2010) proposed a URL recommender system for Twitter users based on social voting and content sources. Generated topics and social interactions were discovered to be significant in the presentation of recommendations. User follow-back recommendations were also proposed by the authors in (Y. Liu, Chen, Li & Wang, 2016),(Takimura, Harakawa, Ogawa & Haseyama, 2018) and (Karidi, Stavrakas & Vassiliou, 2018).

Liang et al., (Liang et al., 2018a) addressed the user profiling problem on Twitter by formulating a dynamic user and word embedding model (DUWE) and a streaming keyword diversification model (SKDM). The embedding model tracked semantic representations of words and users in the same semantic space for similarity measurement. In addition, the authors proposed a streaming keyword diversification model to characterise user profiles over time with top-K keywords.

Enrichment of tweets in construction of semantic profiles using Open Calais [10] ontology for detection of 39 different types of entities, e.g., persons, events, products, etc., was also suggested by Gao et al., (Abel et al., 2011a).

2.4.4 Twitter in Topical Recommendation Systems

Twitter is instrumental in the dissemination of content in diverse domains. Users on the platform relate better to content that mirrors their timely interests based on the content volatility on the platform. Hashtags' recommendations is one such area in short-text microblogs like Twitter. Figueiredo et al., (Figueiredo & Jorge, 2019) proposed a Topic Relevant Hashtag Identification (TORHID) for the retrieval and identification of hashtags, relevance to certain topics on Twitter. The authors made use of a seed hashtag and further classifications to remove hashtags with less relevance. This resulted in more relevant and related hashtags that could be used to deepen the initial search.

An improved Twitter-LDA model with the assumption that topics and background words differ per user was proposed by Sasaki et al., (Sasaki, Yoshikawa & Furuhashi, 2014). The model was further improved considering the time sequence and capability of online inference based on Topic Tracking Model (TTM) (Iwata, Watanabe, Yamada & Ueda, 2009), an LDA based probabilistic consumer behaviour model for tracking user interests.

A hashtag recommender system based on semantic vector representation of tweets is another approach for content recommendations in short-text microblogs (Ben-Lhachemi et al., 2018). Authors used a pre-trained Google news based embedding to represent tweets by averaging per tweet vectors. The features were then clustered using density-based spatial clustering of applications with noise to eventually extract the most similar tweets and recommendations of top-K suitable hashtags in relation to the extracted

[10] `https://rapidapi.com/`

cluster centroids. As much as the model performed well, a correlation between the language in Google news and word-level embeddings were practically not very relevant to the language and word structures in tweets.

On the other hand, Liu et al., (P. Liu, Zhang & Gulla, 2019) proposed a dynamic graph-based embedding model for recommendations of relevant temporal texts and users. The authors modelled a heterogeneous user-item relation that evolved as content changed. The model captured temporal relationships and related texts by embedding their representation in a low dimensional space.

2.5 Multi-Interest Modelling

Short-text microblog users consume content on the platforms based on their interest levels in certain topical content. This interest may either be self-defined or computed via recommender engines which ultimately defines the user interest profiles. They are subdivided as below based on their formulation methodology:

2.5.1 User-Interest Based Profiling (UIP)

Works related to modelling taste profiles based on user-representative interests (User Interest Profiling(UIP)) were pertinent. Banerjee et al., (Banerjee et al., 2009) statistically modelled for the discovery and distribution of user interests in categories such as “games", “food” and “movies”. This was in tweets spanning 10 cities worldwide for a period of four weeks using related keywords. Bao et al.,(Bao, Li, Liao, Song & Gao, 2013) constructed a temporal and social probabilistic matrix factorization model to predict potential user interests in short-text microblogs. Authors in (Garcia Esparza, O’Mahony & Smyth, 2013) constructed a user profiling model based on topical categorization of URLs in tweets. A mean profile prediction accuracy of 0.73 for 32 users over 18 coarse-grained interest categories was achieved. Semantic relatedness for tag

clustering to construct strong user interest profiles (UIP) was proposed by (Goel & Kumar, 2018). Other tags for inferring user interests, such as comments and reviews, were neglected in this work. Nguyen et al., (Nguyen, Sriboonchitta & Huynh, 2017) used soft ratings to model subjective, qualitative, and imperfect information about user preferences. The same approach was used more realistically to express their preferences on products and services.

Hierarchical semantics of concepts from tweets were also used to infer richer user interests expressed as a hierarchical interest graph (Kapanipathi, Jain, Venkataramani & Sheth, 2014a). On the other hand, (Karatay & Karagoz, 2015) proposed a Named Entity Recognition (NER) model for modelling interests for users on Twitter. Zhue et al., (Zhu, Zhou, Deng & Wang, 2019) modelled a user interest graph represented by a hierarchical tree structure covering 167 nodes on three levels. The study also considered interest decay over time. In the same way, authors in (J. Zheng, Wang, Li & Zhang, 2019) developed a hierarchical interest overlapping community (HIOC) detection method by studying similarities in relationships between user profiles, and further presented personalized recommendation models. Lian et al., (Liang, Zhang, Ren & Kanoulas, 2018b) proposed the use of embeddings to jointly model users and their representative content in the same semantic space. The essence was to measure semantic similarity between users and words in inferring user representations.

2.5.2 Social Network-Based Profiling

User connections (bi-directional or unidirectional) in social networks provide a basis for user interest extraction. The friend/follower-followee social connections are modelled as graphs where nodes represent users, and edges, the connections (Abbasi, Tang & Liu, 2014). Jinpeng et al., (J. Wang et al., 2014), proposed a regularisation framework based on the relation bipartite graph. Under this graph, node similarities are evaluated

based on the local link structures instead of explicit links between two nodes, e.g., retweet relations. In social relations, the theory of *homophily* is key. This is basically the tendency of users with common interests to follow each other (McPherson et al., 2001). Pennacchiotti et al., (Pennacchiotti et al., 2012) extracted user interests from the neighbouring users' tweets. The same approach was applied by Mislove et al., (Mislove, Viswanath, Gummadi & Druschel, 2010) in inferencing missing user information based on the one disseminated by neighbours.

However, the theory was extended by assuming that if two users shared followers, then there was a likelihood to also share interests. This meant that local link structures between two nodes indicated node similarity. This to a large extent worked on the assumption that no implicit negative links existed. If they did, then inaccurate user interests were bound to be inferred.

2.5.3 Temporal-Based User Profiling

Microblogs and more so short-text ones are temporal in nature. Their temporality is augmented by the rate of information *gain* and *decay*. Variations in these dissemination patterns are triggered by changes in user interests over time, and thus can be factored in the modelling process (Liang et al., 2018a).

The level of interest in the consumption of certain content defines a user's profile. In short-text microblogs, users tend to define their initial profile information extrinsically by stating their interests, such as when they create new accounts. Normally, users will select from a list of pre-determined interests at sign up time in curation of third-party content bound to be served to them over time. However, such interest(s) may vary over time at the user level. In addition, the selected interest may not be of exclusive importance to them, as they may for example never disseminate any semantically relevant content to the declared interests. This is the main reason why extraction of

interests from the disseminated content and patterns was considered.

Research in modelling user interests more so, in short-text microblogs, has been on going. Ramasamy et al., (Ramasamy, Venkateswaran & Madhow, 2013) proposed a probabilistic way of mining user interests on Twitter factoring tweet times. This was possible using known timings of external events associated with the interests. The approach assumes that a fan is likely to tweet more for example during games in a sports event that at other times. User interests too tend to change over time and are mostly influenced by public trends (Abel et al., 2011a). Here, user interests were modelled at each period as a set of weighted concepts that are either entities or hashtags extracted from user's tweets within the period. Tweets with a shorter temporal distance to the given period are considered more important. Generally, researchers in this sphere considered that temporal dynamics of user interests can improve, e.g., personalised recommender systems.

Jiang et al., (Jiang & Sha, 2015) developed a framework for extraction of user interest changes over time on Twitter. The authors incorporated external knowledge sources to generate more representative user interests. On the other hand, Zhu et al., made use of a hierarchical tree structure factoring in interest decay over time (Zhu et al., 2019). Cami et al., (Cami, Hassanpour & Mashayekhi, 2019), modelled user preferences using a Dirichlet Process Mixture Model where they considered user interactions to construct an evolving Bayesian non-parametric framework. The model performed well in prediction of user preferences and behaviour. In addition, a temporal preference model for detection of changes in the user network structure based on node centrality change events on Twitter and Jam social music datasets was also proposed (Pereira, Gama, de Amo & Oliveira, 2018). A framework for profiling users based on their posting activities more so, on their posting frequency and temporal patterns was also proposed (Ying, Chiu, Venkatramanan & Zhang, 2018).

Figure 2.7 depicts temporality in the changes of topics across two users as identified

by authors in (Ahmed, Low, Aly, Josifovski & Smola, 2011). User *A* is more interested in content related to "Dating" more than other topics in the first few days. However, "Baseball" interest supersedes the other interests on the 30th day. Interest in "Health" and "Celebrity" doesn't change much across the timestamps for this user. On the other hand, user *B* depicts the greatest interest in "Baseball" across the timestamps. Interest in "Jobs" peaks on the 20th day while "Dating" remains fairly low across the time period.

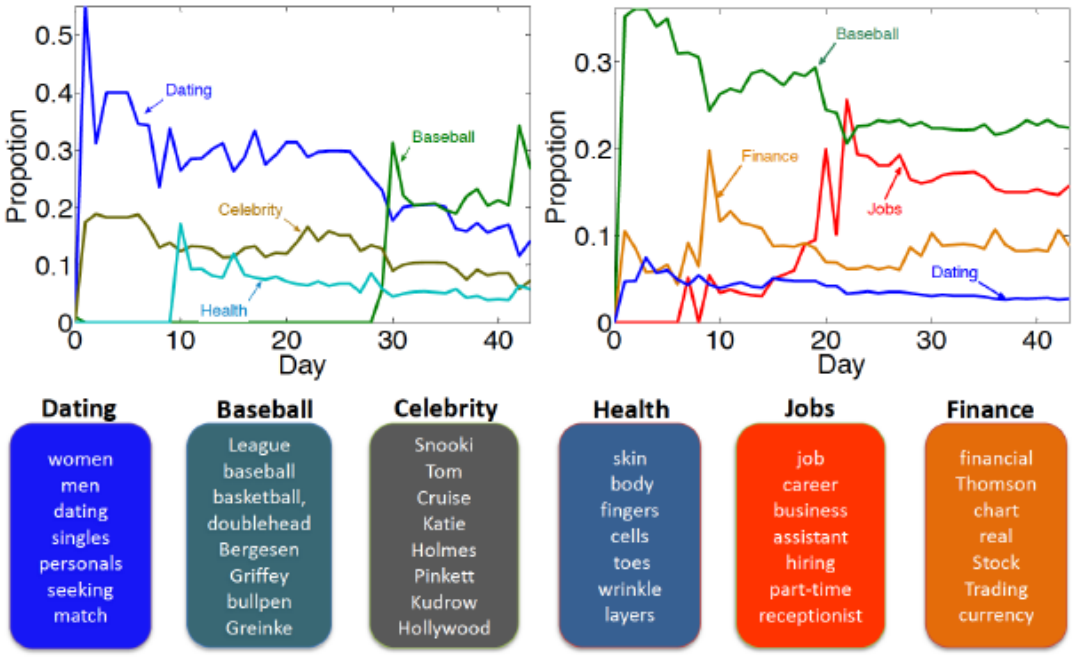


Figure 2.7: Dynamicity of interests in two users over time (Ahmed et al., 2011)

To incorporate temporality of user interests in user modelling strategies, two approaches have been proposed (i) sliding window, and (ii) decay function (Piao & Breslin, 2018). The **sliding window** approach is defined by either the number of items, e.g., the last 20 tweets or the *time period*. This can either be a count of weeks, days or even months. Research work on this domain has been done more so in capturing user interests over time and eventually building user interest profiles at different intervals, simulating the change. For example, Khater et al., (Khater, Elmongui & Gracanin,

2014) captured dynamism in user interests by modelling their daily interests over the extracted topics on Twitter. Yin et al., (Yin et al., 2015a) divided time into multiple time intervals at predefined granularity. The result was a probabilistic model that extracted user-representative topics as interests of users and time-based topics. Elmongui et al., (Elmongui et al., 2015) proposed a tweet recommendation system where user interest profiles were modelled as time-variant topics. A sliding window was applied in calculating the degree of interest in the topics.

With **decay functions**, the main idea related to the replacement/decaying old interests over time. Here, the relevance of each interest was weighted based on its age. Therefore, newer interests had higher weight compared to old ones. Accuracy was in the presentation of user interests where an exponential time-decay function in computation of user interest weights was considered (Orlandi, Breslin & Passant, 2012). Authors in this user study showed that a slower decay function presented a more complete user profile. Abel et al., (Abel et al., 2011a; Abel, Gao, Houben & Tao, 2011b) identified that user interests over time on Twitter are a direct influence of public trends on the platform. The authors proposed a time-sensitive interest decay function based on the distance between the concept occurrence time and a given timestamp to compute the weight of the concept. Ahmed et al., (Ahmed et al., 2011) explored a temporal model with three abstraction levels: (i) user interests over the user's history; (ii) last month's user interest; and (iii) user interest for the past week. The author's argument in the introduction of the three levels was based on the sparsity in user's history. An exponential decay function in this case would not have factored well the long-term interests.

Incorporation of decay functions proved useful for Piao et al., (Piao & Breslin, 2016a). They incorporated user interest dynamics in link predictions on Twitter. Use of decay functions in building long-term profiles improved on the link prediction results. This was in comparison to profiles where the decay function was not considered. The same approach was considered by Nishioka et al., (Nishioka & Scherp, 2016) albeit in

publications' recommendations.

Therefore, building user-representative taste profiles in short-text microblogs involves incorporation of user interests, time and the follower-followee network. With respect to this, several approaches have been proposed. They are outlined in Section 2.6.

2.6 Taste Profiling in Microblogs

Construction of a user's *taste profile* involves extraction of the *user's model* which is largely dependent on the specific user's or group interests. According to Piao et al., (Piao & Breslin, 2018), a *user model* is defined as a data structure that is representative of a user's characteristics. A *user profile* is further defined as the actual representation of the user representative model. Figure 2.8 presents the overall user modelling and profiling process. This is a modification of the profiling architecture, proposed by authors in (Gauch, Speretta, Chandramouli & Micarelli, 2007; Abdel-Hafez & Xu, 2013).

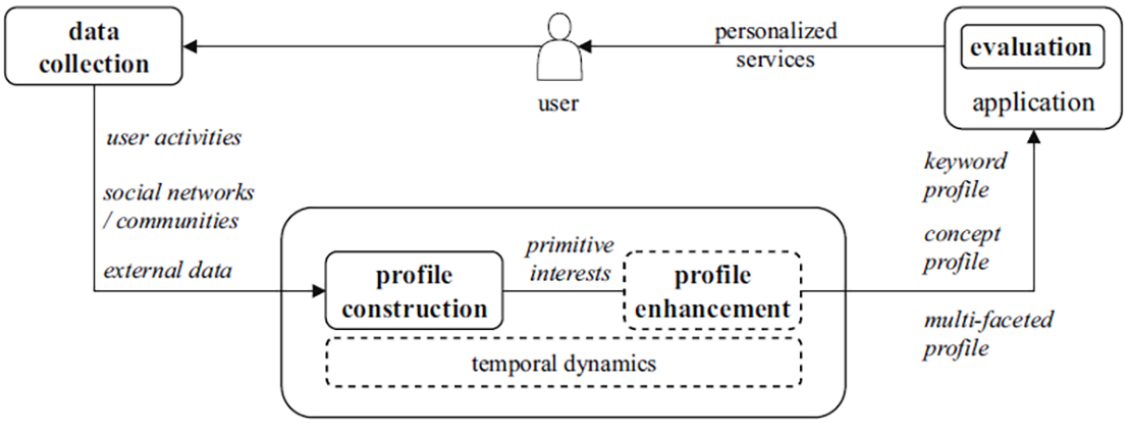


Figure 2.8: User profiling process
(Piao & Breslin, 2018)

As depicted in Figure 2.8, (Piao & Breslin, 2018) detailed the profiling process to

involve modules related to (i) data collection in form of user activities, social networks or external data from ontologies, and (ii) primitive interests extraction. This was either done directly from the dataset or could be enhanced via external Knowledge Bases (KBs). The output was either a keyword-based or a concept-based profile which could then be evaluated for use by recommender systems.

2.6.1 Taste Profiles Representation

Different approaches have been used in varied ways to represent user taste profiles. Taste profiles correlate to user representative content that users consume. As mentioned in Section 2.5.3, the consumption could either be at a specific time or over a time period, depending on the user modelling strategy. User taste profiles in OSNs are broadly categorised into (i) keyword based (ii) concept based and (iii) multi-faceted based ones as introduced by Gauch et al., (Gauch et al., 2007).

Keyword-Based Profiles

In this profiling approach *keywords* or *groups of keywords* are used to represent user interests. Weighting of each keyword is computed to infer the importance of the keyword in the profiling process. Techniques such as the TF·IDF (Term Frequency·Inverse Document Frequency) from information retrieval suffice in building keyword based profiles (Salton & Buckley, 1988; Alsaeedi, 2020).

This approach has been used by researchers in the profiling process in OSNs. Vectors of weighted keywords from tweets and user lists membership descriptions were used in formulating taste profiles (Bhattacharya, Zafar, Ganguly, Ghosh & Gummadi, 2014; J. Chen et al., 2010; Paul, Khattar, Kumaraguru, Gupta & Chopra, 2019). Hashtags [11] in tweets have also been used to infer user interests in the formulation of taste

[11] https://en.wikipedia.org/wiki/Hashtag

profiles. Keywords in tweets were leveraged to represent taste profiles from the extracted keywords (Hannon, McCarthy, O'Mahony & Smyth, 2012; Chowdhury, Caragea & Caragea, 2020; Xu & Zhou, 2020; Wei et al., 2019). Abel et al., (Abel et al., 2011a; Abel, Gao, Houben & Tao, 2011c) also researched on the formulation of hashtag-based interest profiles from tweets. In addition, extracted keywords in tweets were also leveraged to represent taste profiles (Cui, Agrawal & Ramnath, 2020; X. Zheng & Sun, 2019).

Probabilistic approaches like LDA have also been used in modelling user taste profiles. With LDA, topic numbers are preset and the dataset is modelled to extract that number of topics. The topics are defined by the keywords in them. Weng et al., (Weng et al., 2010b) extracted topics from tweets via LDA and represented each tweeter as a probability distribution over the extracted topics. On the other hand, "Twitter-LDA," by Zhao et al., (W. X. Zhao, Jiang, Weng et al., 2011) considered each tweet as a single topic in its topic modelling process.However, the nature of tweets is that they are noisy and short. This made it difficult to extract meaningful keywords (Liao et al., 2012).

Concept Profiles

With limitations posed by keyword-based profiles, researchers proposed a *concept-based* approach to address the shortcomings. The advantage of using concepts from KBs e.g. Wikipedia is that they provided background knowledge for better extraction of concepts. Concepts have been used for varied purposes in user modelling. Their usage as KBs range from simple to complex taxonomies as presented by authors in (Kang & Lee, 2017; Dooley & Božić, 2019; Piao & Breslin, 2016a; Nishioka & Scherp, 2016).

Extraction of hierarchical semantic concepts from tweets for better profiling has also been experimented. Here, interests were presented as a hierarchical interest graph (Yu, Xu, Wang & Ni, 2019; yeon Sung & Kim, 2020). Tommaso et. al, proposed Wiki-mid that made use of Wikipedia in deriving better user interests in multilingual tweets. The

authors used services such as Spotify to extract user preferences (Di Tommaso, Faralli, Stilo & Velardi, 2018). Zheng et. al, came up with the Hierarchical Interest Overlapping Community (HIOC). In the approach, the authors computed user profiles relationships and further presented them as a personalized recommendation model (J. Zheng et al., 2019).

Hybrid Profiles

Fusion of several aspects of interest about a target user in the modelling process results in a hybrid profile. The assumption in this process is that the different aspects complement each other and improve on the extracted profile. Hybrid profiling in detection of fake news is one area that has been explored (Hamdi, Slimi, Bounhas & Slimani, 2020). Hamdi el. al evaluated the credibility of information sources on Twitter using node2vec where features from twitter followers/followees graph are extracted. In the hybrid approach, both user characteristics and social graph are considered in profiling of fake information sources. A fusion of Big Five personality traits and dynamic interests on Twitter has also been leveraged in the profiling of users via a graph-based representation model (Dhelim, Aung & Ning, 2020). A dynamic representation of users' interests and relaxation of the bag-of-words assumption is adopted in the modelling process. Identification of malicious profiles on Twitter is another area where a mixture of important features is used in identifying such users. Sahoo et. al (Sahoo & Gupta, 2019) made use of Petri net structure to analyze user profiles and extract features to train classifiers for prediction of malicious and legitimate users. Making use of concept-based profiles from followee's biographies and list memberships for target users, followers, and followees on Twitter also improved the quality of the URL recommendations (Piao & Breslin, 2017b).

The above aspects correlate to the methodologies in the actual taste profile construction process outlined in Section 2.6.2.

2.6.2 Taste Profile Construction and Augmentation

Literature on the extraction and representation of user interests from collected datasets has been covered in Sections 2.4 and 2.2.1. Details on works related to the actual construction and augmentation of taste profiles are detailed this section.

Taste Profile Construction Approaches

A *profile constructor* determines the importance user interests by computing their weights with respect to the specific user. A few weighting schemes have been proposed by researchers based broadly on the following criteria. A *weighting scheme* is a function that determines the importance of the user interests.

(i) Empirical Approaches - A common approach in this weighting scheme is the Term Frequency (TF). Here, the importance of a word relative to the user is determined by the count of the term in the data source. Researchers in (Abel et al., 2011a; Kapanipathi et al., 2014b) utilised this approach, despite its simplicity, with interests, as concepts with scores would account for the weighting scheme. However, this approach favoured commonly occurring words. Thus common words to a user that may not be of interest to the same user, ended up with higher weights. IDF (Inverse Document Frequency) was introduced to complement TF. The IDF score of y with respect to a user x on x's tweets was measured as $IDF_x(y) = log[\frac{allusers}{usersusingyatleastonce}]$ (J. Chen et al., 2010). Several variants of IDF usage have been applied by researchers in interest extraction and taste profiles enhancement (Nishioka & Scherp, 2016; Gao, Abel, Houben & Tao, 2011; Piao & Breslin, 2016b). Modified versions of the same as well as a comparison with other weighting approaches like Text-Rank have also been used in ranking users interests (Mihalcea & Tarau, 2004; Vu & Perez, 2013).

The same approaches have been modified to fit OSNs where an usage of the user network structure in the construction of user taste profiles was explored (Lu et al., 2012;

J. Chen et al., 2010). A set of high-interest keywords for followees were extracted to model a user's profile using followees' tweets. The same approach, albeit with list memberships of followees, was followed by other researchers (Piao & Breslin, 2017b; Bhattacharya et al., 2014).

(ii) Probabilistic Approaches - In using the empirical approaches as described above, the assumption is that user interests are explicitly dependent on keywords in the text. Probabilistically, a similarity measure between a post and an entity would be used to infer user interests. The Explicit Semantic Analysis (ESA) algorithm(Gabrilovich, Markovitch et al., 2007) was used by authors in (Lu et al., 2012; Narducci, Musto, Semeraro, Lops & De Gemmis, 2013) in the computation of the similarity between texts, to generate weights for each user.

User interest weights have been modelled in unsupervised ways by researchers. Twitter-LDA is one approach where Weng et al., (Weng et al., 2010b) aggregated all tweets per user in one long document and applied LDA on the long documents. LDA was also used in the inference of topic distributions for each user where DBPedia entities were set as topics (Zarrinkalam, Fani, Bagheri & Kahani, 2017; Trikha, Zarrinkalam & Bagheri, 2018; Zarrinkalam, Fani, Bagheri, Kahani & Du, 2015). A probabilistic generative model was also proposed by Budak et al., (Budak et al., 2014) to extract user interest profiles as probabilities over ODP (Open Directory Project [12]) categories. Target user's posts, user content dissemination patterns and the friendship network influence were considered in the modelling process. Another probabilistic framework for inferring user interest profiles was proposed by Sang et al., (Sang, Lu & Xu, 2015).

Taste Profile Augmentation

User interests from disseminated content can be further improved using external knowledge to deliver better user-representative taste profiles. From the literature , this

[12] https://en.wikipedia.org/wiki/DMOZ

improvement was leveraged in the following ways:

(i) Making use of Hierarchical knowledge - Based on hierarchical KBs such as Wikipedia categories, taste profiles could be inferred (Kapanipathi et al., 2014b). In the refinement process, Wikipedia entities were extracted from tweets as initial interests. The entities were then used as activated nodes for applying a spreading activation function (Collins & Loftus, 1975) on the hierarchical KB to infer weighted categories to represent taste profiles. Work by Kapanipathi et al., (Kapanipathi et al., 2014b) on the spreading function was adapted by other researchers as long as the hierarchical KB was available (Nishioka & Scherp, 2016; Piao & Breslin, 2017a; Große-Bölting, Nishioka & Scherp, 2015; Besel, Schlötterer & Granitzer, n.d.). Nishioka et al., (Nishioka & Scherp, 2016) extracted entities from the economics domain using the spreading function. The same was replicated in the computer science domain using the ACM CCS concept taxonomy (Große-Bölting et al., 2015). Similarly, Besel et al., (Besel et al., n.d.) used WiBi (Flati, Vannella, Pasini & Navigli, 2014) in mapping followees' Twitter accounts to Wikipedia categories in the entity extraction process. Twixonomy methodology, a Wikipedia based category taxonomy, was proposed by Faralli et al., (Faralli, Stilo & Velardi, 2017) as a more accurate taxonomy. It is a graph pruning approach based on a variant of Edmonds optimal branching (Dantzig & Veinott, 1968).

On the other hand, other KBs have been explored in the augmentation of profiles based on primitive interests. News items have been leveraged as external KBs in this perspective (Kang & Lee, 2017). The authors proposed a process of mapping news categories from two news portals[13] to tweets. Furthermore, they used Wikipedia to vectorise tweets and news categories in the same semantic space to bridge the semantic gap between terms in tweets and news articles. The same approach, but with other KBs such as DBPedia, Freebase, and Yago, have been used by other researchers(Suchanek, Kasneci & Weikum, 2007; Bollacker, Evans, Paritosh, Sturge & Taylor, 2008; Jiang

[13] `http://news.naver.com/,http://news.nate.com//`

& Sha, 2015). Bhargava et al., (Bhargava, Brdiczka & Roberts, 2015) made use of Facebook Pages and Yelp[14] categories to extract features such as hashtags and document categories in leveraging user interests.

(ii) Graph-based knowledge - In addition to hierarchical approaches, researchers have made use of graph-based approaches in augmenting the generated user profiles. Knowledge Graphs (KGs) and instances of the classes in the ontology have been investigated (Färber, Ell, Menne & Rettinger, 2015). KGs in essence provide related entity information in addition to the high-level categories. Abel et al., (Abel, Hauff, Houben & Tao, 2012) used DBPedia's background knowledge in the augmentation of user taste profiles. The same approach was adapted in enriching primitive interest categories with DBPedia categories in profiling tweeters (Peñas et al., 2013; Orlandi et al., 2012). Piao and Breslin (Piao & Breslin, 2018) made use of three property-based propagation strategies, i.e., class, category and property, on DBPedia in generating more precise twitter URL recommendations.

Alternatively, Wikipedia entity graph, as opposed to DBpedia graph, entailed the mention of other entities as opposed to predefined ontological properties. Wikipedia entity graph was used in enhancing entity-based primitive interests (Lu et al., 2012). The intuition was that a user with interest in, for example, *Samsung Tablet*, may ideally be interested in other *Samsung* products, as compared to general *Android mobile devices*. Entities were extracted from tweets as interests and were expanded on the Wikipedia entity graph via a random walk methodology.

(iii) Collective Knowledge - Augmenting taste profiles with other related interests is another approach in taste profiling. The collective term for this process is Frequent Pattern Mining (FPM). The idea is to find the frequency in sets of items that co-occur. This approach has been used in identifying topics that frequently co-occur (Faralli, Stilo & Velardi, 2015; Trikha et al., 2018) in augmenting the taste profiles generation

[14] `https://www.yelp.com/`

process. The authors used FP-Growth algorithm in mining pattern frequencies (Han & Pei, 2000).

2.7 Chapter Summary

Several relevant topics and related literature were reviewed in this section. Research works related to OSNs and specifically Twitter were discussed. This was followed by approaches in knowledge extraction in short texts. User interests were identified to be building blocks for any taste profiling model. A few approaches in the semantic user interests extraction process were discussed. The consensus is that their variations are in terms of the dataset nature as well as task at hand. In the last part of this chapter, research work related to taste profiles construction and augmentation were discussed. Ideally, profiles are products of a modelling process involving the extracted user interests.

Chapter 3 introduces one of the approaches in modelling and augmentation of short texts. The rest of the chapters contribute largely to the user interests extraction/augmentation and overall taste profiling processes. Each chapter has an experimental and results discussion section.

Chapter 3

Metamodel LDA (MELDA)

3.1 Introduction

In this chapter, the first approach in modelling short texts as part of the user profiling process is presented. This is further refined in Chapter 4 where user interests are extracted for further taste profiling. Interest in the extraction of coherent topics in short, noisy, and topically constrained data characterises the MELDA approach. This type of data has limitations in word co-occurrence, and thus a sparse vocabulary. Therefore, it is not a trivial task even for topic modelling algorithms to derive interpretable tasks from such data (Chang, Gerrish, Wang, Boyd-Graber & Blei, 2009).

Comprehending social media discussions in short-text microblogs is fundamental in knowledge-based applications such as recommender systems. Twitter, for example, provides rich real-time information in keeping up with its streaming nature. Making sense of such data without automated support is not feasible due to its vast size and nature. The problem becomes even more complex when the data in question have a low variance in terms of topical diversity. Therefore, an automatic method for understanding textual patterns in such topically constrained data needs to be developed. A major challenge to building such a system is in its ability to comprehend the nature of the data

regarding diversity of word structure correlations, vocabulary sparsity and distinguishing factors in the generated topics. In this chapter, a novel semi-supervised approach called Metamodel Enabled Latent Dirichlet Allocation (MELDA) is presented to address this challenge. Compared to state-of-the-art approaches, the model incorporates a domain-specific metamodel. The *metamodel* is defined as a set of topic label vectors derived from long texts to guide the learning process in shorter texts.

3.1.1 Knowledge Extraction in Short-Text

Shorter texts are more difficult to mine and analyse due to the limited vocabulary compared to standard long-text datasets. Slang, different languages of expression, and inconsistent and sparse vocabulary across tweeters present formidable challenges to extracting knowledge in such data. Events on short-text microblogs generated through Twitter give rise to topics of discussion in the form of trending topics. A *topic* is a collection of words or phrases that refer to a popular but temporal concept. For example, Twitter and Facebook provide a real-time list of trending content and topics for users. This includes posts from friends, discussions outside their circle, as well as breaking news, etc.

In addressing the short texts mining challenge, the aim is to extract more coherent topics in such short texts by incorporating domain-specific long texts to augment the mining process. A novel semi-supervised learning-based approach called Metamodel Enabled Latent Dirichlet Association (MELDA) is presented to solve this problem. The approach is semi-supervised owing to its modeling process. The modelling process is guided by incorporating a metamodel built from an external but semantically relevant long-text dataset. Taking the metamodel into account refines the topic modelling process for short texts and enables better topic coherency, as shown in experimental results in Section 3.4.2. The metamodel is important in the LDA initialisation phase where words

and their respective co-occurrence patterns bias word distribution over the short-text dataset. A *seed confidence value* that defines the percentage of topics convergence towards certain words in the seed topic set from the metamodel is introduced.

In the evaluation of MELDA, metamodel topic label vectors are generated from a collection of smartphone reviews and distributed over product support related tweets (short-text dataset). This allowed for generation of more interpretable topics as detailed in Section 3.4.2. The dataset of choice is termed as *topically constrained* as it has low topic diversity. In the smartphone dataset that was experimented on, word co-occurrence in certain aspects was high. For example, complaints related to the camera aspect were almost similar to those of the screen or battery aspects. Therefore, a topically constrained dataset contrasts conventional datasets related to, e.g., news, sports or politics that are linguistically diverse and with well-defined topics. The following contributions were made in this section :-

1. A two-step process is proposed to incorporate long-text external knowledge in the metamodel for modelling topics in shorter texts:-

   (a) **Extraction of topic labels** - Topic labels from the metamodel are extracted via LDA. The assumption is that LDA works well on long texts (T. Wang, Viswanath & Chen, 2015; Griffiths & Steyvers, 2004; Blei et al., 2003a). The metamodel is derived from word co-occurrence patterns and is used to guide the initialisation of topics in shorter texts.

   (b) **Distribution of short texts terms over the metamodel** - Tweet terms are distributed over the metamodel topic labels at the initialisation step resulting in a *metamodel topic labels-tweets matrix*. The novelty in the semi-supervision aspect lies in the distribution of topic label vectors derived from seed topics over short texts to anchor the initialisation phase. A *seed confidence value* is introduced to determine the level of restraint which ultimately

determines the topical variance between the two sets as well as the term co-occurrence based topic modelling bit. The guidance and seeding process in modelling aims to generate more coherent topics. This is not the case in vanilla LDA (Blei et al., 2003a) and its variants (W. X. Zhao, Jiang, He et al., 2011; W. X. Zhao, Jiang, Weng et al., 2011; Z. Liu et al., 2010; Weng et al., 2010a; Andrzejewski et al., 2009; Z. Chen et al., 2013b; X. Wang et al., 2011).

2. **Topically constrained dataset** - Regarding the dataset, a topically constrained short-text dataset is presented and modelled. This is a dataset that has low topic variance in terms of overlapping topics and word co-occurrence patterns. Unlike conventional long and short-text datasets, it lacks well defined categories and vocabulary variance is not overly diverse.

3. **Evaluation against baselines** - An evaluation of our approach against state-of the art baseline approaches namely LDA (Blei, Ng & Jordan, 2003b), Twitter-LDA (Weng et al., 2010b), and SILDA (He, Wang & Jiang, 2017a) on the same noisy dataset with the same parameter settings is performed. A quantitative evaluation of our results demonstrates better semantic interpretability of the modelled topics.

3.1.2 Notations

The below notations and respective descriptions are used in this chapter: -

θ - LDA's topic-word multinomial distribution.

β, δ and γ - LDA and MELDA related Dirichlet Priors.

T - Collection of tweets with each tweet t having N_t words.

M - Number of documents in a MELDA modelled dataset.

K- Number of topics in MELDA and LDA.

D - Number of seed topic-sets from the metamodel.

ψ - Represents the integration of the metamodel topic label vectors with raw tweet vectors forming a tweet-metamodel topic label preference matrix (DXT).

3.2 Background and Problem Statement

3.2.1 Latent Dirichlet Allocation (LDA)

LDA is an unsupervised approach in textual data knowledge discovery. It generates a summary of pre-set topics through a discrete probability distribution over words. Per-document distribution over the generated topics is then inferred. Each document is therefore interpreted as a mixture of various topics with the topic distribution assumed to have a sparse Dirichlet prior. This sparsity ensures that documents only cover a small set of topics and that topics can also be captured by a small set of words that reduce ambiguity in the generated set of topics (Blei & Lafferty, 2006). Formulation of a topic in LDA is based on term co-occurrence likelihood. For example, a term may be presented in different topics, but its surrounding words define the interpretability of the topic. In our case metamodel seed topic vectors are distributed over raw tweets to generate more interpretable topics.

LDA Generative Process

In LDA's topic generative process as shown in Algorithm 1, θ and β are the models' hyper-parameters, x as the document/tweet with d_x being words/terms in corpus D. θ_d is the topic-word multinomial distribution with parameter vector α (Blei et al., 2003b). The second loop presents a multinomial distribution of topics z_w over words w_w with β as the Dirichlet Prior.

Algorithm 1 LDA Generative Process

1: **for** Document x with d_x words in Corpus D **do**
2: Randomly choose $\theta_d \sim (\alpha)$
3: **for** position w in d_x **do**
4: Choose a topic $z_w \sim (\theta_d)$
5: Choose a word w_w from $p(w_w|z_w, \beta)$
6: **end for**
7: **end for**

3.2.2 Problem Statement

Syntactic rules are not defined in online content dissemination more so in short-text microblogs. Moreover, data sources like Twitter present short and noisy texts that make it hard for not only machines but even humans to interpret the underlying knowledge. One challenge with such data is the short document lengths and term co-occurrence sparsity. A single tweet comprises of 140 characters (recently extended to 280) which poses a comprehension challenge as textual understanding is based on word co-occurrence, that is limited in this case. Long-text modelling algorithms have proved not to work well on such texts due to this limitation particularity when fully unsupervised (Chang et al., 2009).

Apart from the challenges that stem from short texts, the dataset even on manual inspection presents a unique challenge. Topic variance is quite limited, which makes topical comprehension more difficult. The dataset as presented in Section 3.4.1 is unique to support related tweets of several smartphone brands. Topics of discussion are quite similar, but some unique variances are also present. Discovery of meaningful patterns in such short texts is therefore not a trivial task especially in topic discrimination.

3.3 Metamodel Enabled LDA (MELDA)

3.3.1 Overview

To address the problem presented in Section 3.2.2, a guided-based modelling process is adopted by introducing a metamodel from a semantically relevant external long-text data source. The approach is named Metamodel Enabled LDA (MELDA) [1]. With semi-supervision, even long-text topic modelling algorithms perform well on short-text documents. The only assumption in this approach is that prior knowledge in the domain of interest is available to guide the shorter texts learning process.

MELDA models each tweet as a mixture of topics, the same way LDA does. In MELDA, each document is considered to have a set of topics that are assigned via LDA. Therefore, each tweet in the dataset is distributed over the metamodel topic-sets referred to as the *seed topic sets*. A topic-word is generated in relation to a conditional distribution between regular topics and *seed topic sets*. This introduces two latent variables: tweets-based topics (regular topics) and seed topic sets (seed topics). A *seed topic* is a non-uniform probability distribution over the words in its set. The model infers the probability distributions of *seed topics sets*. Topic generation in MELDA is two-fold: (a) generation of topics via the *seed topics*, (b) generation of regular topic sets from short texts known as *regular topics*. There is a restriction on the seed topics distribution to only generate words from the seed set. Regular topic distribution on the other hand is not restrictive and can generate any word, including the ones in the seed topic sets.

MELDA differs from conventional LDA in its ability to control how words are assigned to topics in the initialisation step. A *seed confidence value* is introduced. This is a value that controls the level of bias that seeded words incline towards the seed topics. With a seed confidence of 0.15, for example, bias is set at 15% more towards

[1] This work has been published in IEEE Access journal.

seed topic sets. The process is iterative which after several iterations helps the model to converge.

For simplicity, a one-to-one correspondence between regular and seeded topics is assumed. A *tweet-metamodel matrix* representing the semantic relevance of a tweet to the metamodel topics is generated in the process. This makes it computationally possible to assign each raw tweet to a specific seed topic set based on the seed confidence value. MELDA adds another latent factor, seed topic set and seed confidence value. Therefore, the probability that a word n arises in topic k is derived as below,

$$p(n|k) = \int_d \int_k p(k|t) \times (d|k) \times p(n|d,k) t d t k t x \tag{3.1}$$

where n is the word, d is the metamodel topic label, t is the tweet, and k is the topic. Relevant words in different topics are therefore correlated to respective topics that they fall under. The model then chooses the right t corresponding to topic k based on a seed confidence value x. This is instrumental in resolving topic ambiguity problems where a word falls in two or more topics as well as accurately modelling under-represented topics.

3.3.2 Design Framework

As mentioned in Section 3.1, the topic modelling process is two-stepped. Firstly, topic labels from a long-text smartphone reviews dataset as the metamodel are extracted. This is primarily for extraction of specific aspects in form of topic labels known as *metamodel* topic label vectors. The topic label vectors are then distributed over the raw tweets. Topics and Term Frequency-Inverse Document Frequency (TF-IDF) of relevant terms in each topic as label vectors in the metamodel-tweets matrix are ranked. For evaluation purposes, three dominant smartphone aspects (*battery, screen, camera*) are selected to guide the supervision process at the initialisation step in the short-text

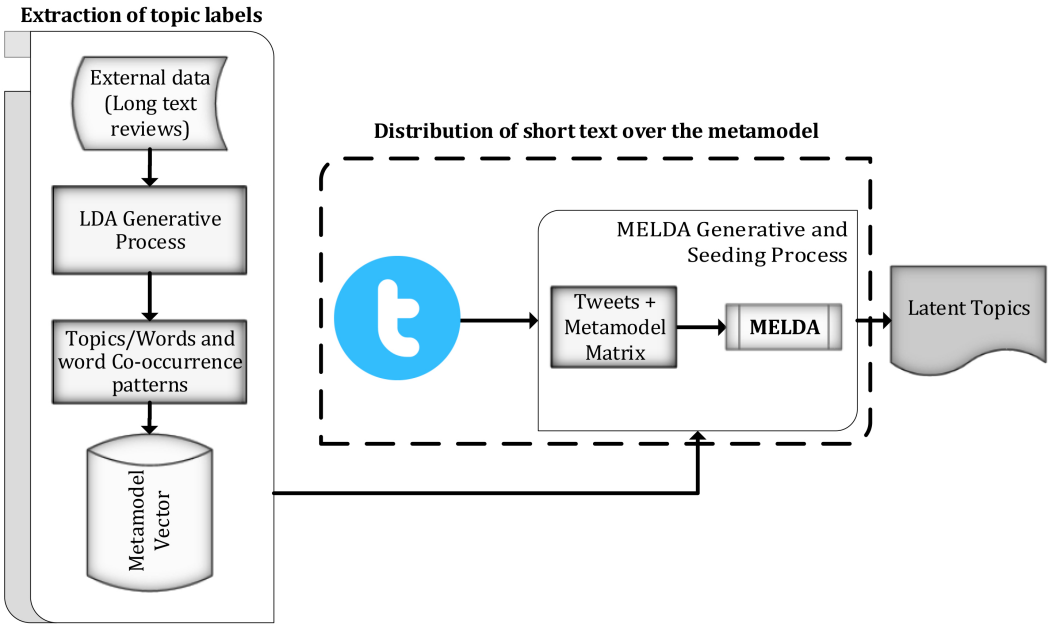


Figure 3.1: MELDA Framework: Pre-processed raw tweets are distributed over topic labels from the metamodel.

dataset. Figure 3.1 illustrates this process. The left-side box labelled "extraction of topic labels" in the figure represents the process of extracting the metamodel vectors (topic labels and relevant word ranks). The associated topics and related words inform the word co-occurrence patterns to be used in initialising topics in the shorter texts. The assumption is that the long texts are contextually relevant to the short texts. Metamodel topic label vectors are then distributed over the short tweets which is MELDA's last step in its generative process. This process is also LDA inspired with bias using a *seed confidence value* to guide the initialisation step. In Figure 3.1, this process is illustrated in the right-side box labelled "distribution of short-text over metamodel topic vectors".

3.3.3 Generative Process

MELDA's generative process is inspired by work presented in (He et al., 2017a; Jagarlamudi, Daumé III & Udupa, 2012) which factored social interests and topical categories respectively as Dirichlet priors for distribution over the dataset. However, MELDA is

Algorithm 2 MELDA Generative Process

Draw $\lambda \sim Dir(\delta)$, $\pi \sim Bernoulli(\tau)$;
for each topic $k \in \{1, \ldots, K\}$ **do**
    Derive the topic distribution over seed topic sets $\varphi_k \sim Dir(\gamma)$
    for each seed topic set $d \in \{1, \ldots, D\}$ **do**
        Derive a per-topic and per seed-topic set distribution over words $\phi_{k,d} \sim Dir(\beta)$
        for each tweet $t \in \{1, \ldots, T\}$ **do**
            Draw $\alpha^{(t)} = \alpha \times \psi_{t,d}$
            Derive a topic distribution per tweet $\theta^{(t)} \sim Dir(\alpha^{(t)})$
            for each word $n_{t,i}$ where $i \in \{1, \ldots, N_t\}$ **do**
                Select an indicator $\zeta_{t,i} \sim Multinomial(\pi)$
                Select a topic $k_{t,i} \sim Multinomial(\theta^{(t)})$
                Select a seed topic set $d_{t,i} \sim Multinomial(\varphi_{k_{t,i}})$
                Derive word $N_{d,i} \in Multinomial(\phi_{k_{t,i},d_{t,i}})$
                if $\zeta_{t,i} = 1$ **then**
                    Select word $n_{t,i} \sim (\lambda)$
                else
                    Select word $N_{t,i} \sim (\phi_{k_{t,i},d_{t,i}})$
                end if
            end for
        end for
    end for
end for

based on seed topic sets generated and ranked from the metamodel. This guides the seeding process especially in short-text word co-occurrence patterns.

In Algorithm 2, topic distribution λ is drawn from a Dirichlet Prior δ just like in the LDA generative process before the first loop. The first loop represents a multinomial distribution over seed topics sets φ_k as well as from a Dirichlet prior γ. A tweet-metamodel topic label preference matrix is introduced in the third loop for revaluation of $\alpha^{(t)}$ in the tweets and metamodel topic sets.

Word-topic and document-topic distributions are also computed in this step. Tweets with higher correlation to the seed topic set after adjustment of the seed confidence value, are assigned a higher hyper-parameter $\theta^{(t)}$ in the fourth loop. Topics $\zeta_{t,i}$, with a higher correlation to the seed topic sets from the metamodel as per the seed confidence value

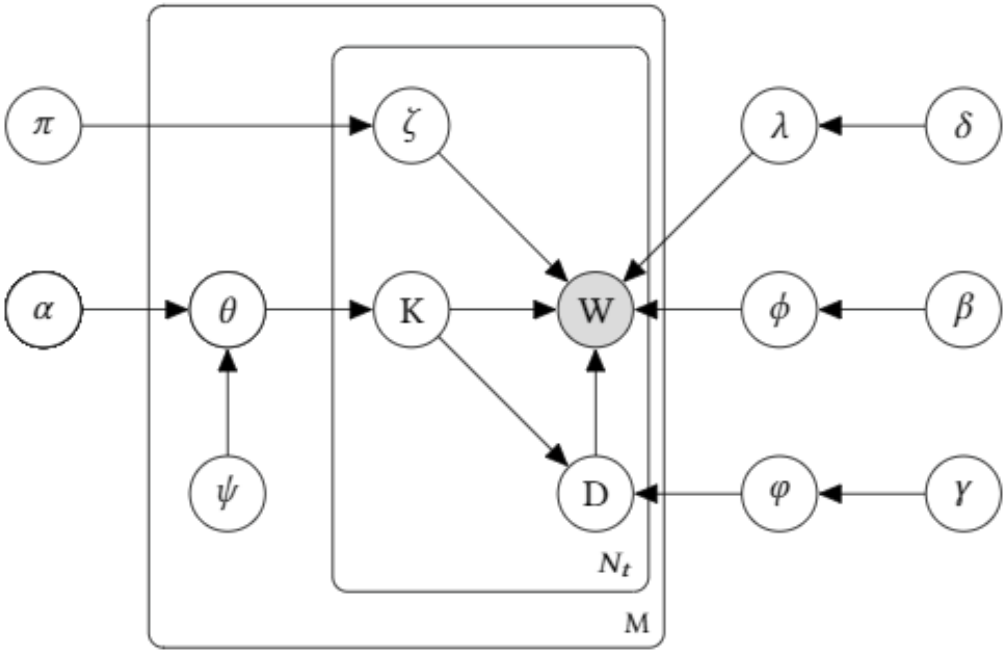


Figure 3.2: MELDA probabilistic graphical model

are modified to include seeded words. This is illustrated in Section 3.3.1. Otherwise, regular topics take precedence. The process is demonstrated in the last two conditional statements in Algorithm 2.

Figure 3.2 depicts the flow of MELDA's probabilistic graphical model. Notations in Algorithm 2 complements the plate notation representation. Two parts in this plate notation emit a word. Topic distribution λ, is drawn from a Dirichlet prior δ. Vocabulary distribution ϕ from the seeded set draws emitted words from a Dirichlet distribution β. This step is illustrated in the second loop of Algorithm 2. π is a multinomial distribution that governs the choice between the seeded set and regular topics. M is the number of documents in the collection, N_t, the number of terms/words in the collection, K as the number of latent topics in the collection and D, the number of seed topic sets.

(Battery Fires): battery, note, samsung, galaxy, fires
(Screen): screen, iphone, apple, new, get, replacement, cracked, one
(Camera): selfie, 16mp, panorama, pixels, zoom, focus
(Battery/Screen/Camera): screen, camera, battery, screenshot, update, draining

(a) Regular topics and respective term distributions

(Battery Fires): battery, note, samsung, galaxy, fires, draining
(Screen): screen, iphone, apple, new, get, replacement, cracked, one
(Camera): selfie, 16mp, panorama, pixels, zoom, focus, camera
(Service): replacements,delay,contact,update

(b) Seeded topics and respective term distributions

Figure 3.3: Sample model output: (a) shows a list of incoherent terms in topics, (b) lists converged topics and respective term distributions. Colour codes differentiated the topics.

3.3.4 Sample Representation

The ***Battery/Screen/Camera*** topic in Figure 3.3 (a) shows several overlapping terms. Terms converge to represent topics and, in this case words like cameras, screen and battery overlap in the *Battery/Screen/Camera* topic. This makes topic discrimination difficult. In a supervised setting, proper labelling or more training can fix the problem. However, this is difficult in a purely unsupervised setting. Such a situation can arise if there are not enough documents about specific topics. In addition, the dataset's topical variance is constrained in addition to a sparse vocabulary which renders topical discrimination difficult. In the approach, terms related to battery, screen, and camera could converge around related topics in the metamodel seed topic sets. At the end of this process, terms converge towards specific topics as in Figure 3.3 (b).

Conventional LDA assigns each word randomly to a topic in the initialisation step. As mentioned in Section 3.3.1, the randomness could either be skewed if the α value is small, and vice versa. Finding out which terms belong to what topic follows in the generative process. For example, finding the topic for the term *Xperia* follows this process. LDA modelling process first assumes that each term in the corpus belongs to a certain random topic. *Xperia* is thus assumed to belong to any of the topics in Figure 3.3 (a). The next computation is pairwise combinations regarding which other terms co-occur with *Xperia*. The most common topic among the co-occurring terms becomes also *Xperia*'s topic.

Algorithm 3 Topic Modelling Initialisation Process

for $x \in X$ **do** ▷ X - all documents
  for $w \in W$ **do** ▷ W - all words
    for $z \in Z$ **do** ▷ Z - all topics
      $w \in z$ = Count w across all documents in z
      $x \in z$ = Count x across all words in z
      Tokens in z = Count(all assignments in z)
    end for
  end for
  $p(z|w,x)$ = (($w \in z$) ×(x $\in z$)) / Tokens $\in z$)
  $w_z = max(p(z|w,x))$
end for

The probability of *Xperia* fitting in topic Z *(Battery Fires, Screen, Camera, Service)* when it occurs follows this procedure. First, the number of *Xperia* word counts assigned to topic Z multiplied by the number of other words already in Z is computed. The product is then divided by the total number of times any word is assigned to Z (Jagarlamudi et al., 2012). Algorithm 3 captures this initialisation process.

Xperia will most likely lie between *screen* and *camera* topics, because contextually, it co-occurred with terms related to camera and screen more than *OnePlus*, *Huawei* or the other models. Manual inspection of the dataset and metadata proves so. Iterations over term co-occurrence patterns eventually lead to model convergence. Our concern is with regard to terms convergence towards certain seed topic sets in the metamodel. An assumption in the MELDA generative process is that external data and its underlying topics must be known beforehand. Word distribution patterns in the metadata set provide clues on the intended word co-occurrence patterns. For example, the LDA generative process might classify *Xperia* in the *screen* topic but that may not be the case in the metadata set. Guidance is therefore needed to fine-tune the model towards assigning *Xperia* to, say, the camera topic or even service, if that is a pertinent issue with *Xperia*. A seed confidence value is also introduced in the initialisation step, to boost the convergence of the *Xperia* term towards the topic of choice as per the seed

topic sets.

Based on the seed topic sets, certain patterns may be exhibited. For example, *replacements* and *contact* in Figure 3.3 (b) may be seeded towards *Battery Fires* topic if that reflects the word co-occurrence pattern in the metamodel seed topics. With more iterations, an increase in the number of seeded words will be noted until the model converges. For seeded documents, the word count for w will be higher thus $p(z|w, x)\alpha$ count across all docs w belongs to z.

3.4 Experiments

A quantitative evaluation of the model is presented in this section. Different authors investigated different aspects and techniques related to modelling of topics in short-text sets. The rationale for choosing baselines as comparatives with our approach is discussed below.

1. LDA (Blei et al., 2003a) is a well proven method of extracting coherent topics, especially in datasets with adequate vocabulary.

2. Twitter-LDA(W. X. Zhao, Jiang, Weng et al., 2011) is a specific modelling approach for tweets. In its generative process, an assumption is made that each document (tweet) represents exactly one topic. It thus reports significant improvements when it comes to modelling topics from short texts.

3. SILDA (He et al., 2017a) introduces the pre-learned interest knowledge from the tweets themselves based on segmented user interests as prior knowledge. The model pre-learned interest-word-sets and the tweet-interest preference matrix are the prior knowledge that is incorporated into the topic modelling process.

3.4.1 Datasets and Settings

Dataset

The core interest in this process is in modelling coherent topics of interest from short texts. Therefore, Twitter provided an ideal testing scenario for mining short-text data. A real-life Twitter dataset related to smartphones was used. Support related tweets for *Apple, Huawei, Samsung, Sony* and *Oneplus* brands were collected to support the modelling process. The phone models under these brands are diverse, which offered good scope for simulating a real-life scenario. The data was collected between January and December 2016 via Twitter's streaming API [2]. Support-related tweets were chosen as they offered deeper insights into specific phone aspects. 490,231 English tweets were collected from 268,465 unique users, not including those disseminated by the brand handles that were mainly generic replies. Manual inspection of the dataset revealed that it was quite compact in terms of topical variance. This implied that topical diversity was small compared to conventional datasets related to, e.g., news, etc. Each entry in the dataset represented a single tweet with its associated metadata, e.g., geo-location, mentions, hashtags etc.

Metamodel Formulation

As described in Section 3.1, the goal was to improve on the learning process of shorter texts for coherency in regular topics. Metamodel formulation from a conventional long-text dataset that was semantically related to the short-text dataset was ideal. Therefore, an assumption in this approach is that prior knowledge on the choice of external data is available.

[2] `https://developer.twitter.com/en.html`

External Data Choice

Contextual and linguistic relevance is paramount in selecting the external data source. **Semantic relevance** and **textual descriptiveness** are ideal indicators in selection of the external data source. In this case, popular phone reviews from Amazon matched this criterion. **65,528** Amazon smartphone brands reviews [3] were modelled as external data. The collected data included the product and brand names, price, rating, review votes and the review content. For purposes of metamodel generation, every other attribute from the dataset was dropped, except for the review content attribute.

Pre-processing

Since the two datasets were syntactically dissimilar, pre-processing them to a machine comprehensible format was done differently. For the long texts, Natural Language Toolkit (NLTK) stop words list [4] was used to filter out common words in the dataset. Collocation detection and lemmatisation was also performed. Moreover, punctuation and numbers were also filtered out.

For the tweets, the process was slightly altered, especially when it came to processing specific aspects of tweets that were missed in metadata. Each tweet in the dataset was processed as a single document. Tweets comprising of four words or fewer were stripped off the dataset as they were likely to introduce noise into the data. Unlike conventional long texts, tweets differ due to the usage of emoticons, hashtags, user mentions, etc. Emoticons were translated to a textual format and added them to the stop-word list for removal. Hashtags were not of preference, and thus were also filtered from the dataset. The interest was in the vocabulary co-occurrence patterns, of which hashtags played a minimal role, justifying their exclusion.

[3] https://www.kaggle.com/PromptCloudHQ/amazon-reviews-unlocked-mobile-phones

[4] http://www.nltk.org

Regular expressions via the regex package [5] were applied to simulate possible URL structures to remove all URLs from the tweets. With most of the noise removed, the resultant clean short-text dataset was tokenised via NLTKs Tweet Tokenizer [6] in preparation for input to the model. A sample tweet before and after pre-processing is shown below: -

Before: *"Samsung to halt Note 7 production temporarily amid reports of battery fires w/ replacement devices http://www. wsj.com/articles/ samsung-to-halt-galaxy-note-7-production-temporarily-147606452 #CJ278 #note7"*

After: *'samsung', 'halt', 'note', 'production', 'temporarily', 'amid', 'reports', 'battery', 'fires', '/', 'replacement', 'devices'*

Parameter Settings and Model Tuning

All models in Section 3.4, metamodel and MELDA were trained using 5000 Gibbs iterations with $\alpha = 50/K$, where K was the number of topics, $\beta = 0.01$. The seed confidence value γ was set at 0.15. This value is basically a user-defined weight depicting the variance between seed and regular topic sets. The metamodel parameters remain the same as the above except for the seed confidence value that, was not applicable. Seed topic sets were generated with $\alpha = 50/K$. Heuristically, setting parameters this way led to optimal topic modelling results as reported by (Griffiths & Steyvers, 2004), (He et al., 2017a). Therefore, the same settings were maintained in the experimental setup.

The same parameter settings were replicated for SILDA (He et al., 2017a) except for the seed confidence value, which is not present in their approach. In Twitter-LDA, parameters settings replicated those of authors with $\beta = 0.01$ and $\gamma = 20$. Seed topic sets and confidence value were not factored in Twitter-LDA (Z. Chen et al., 2013b). It

[5] `https://pypi.python.org/pypi/regex/`
[6] `http://www.nltk.org/api/nltk.tokenize.html`

was difficult to set the optimal number of topics, thus trials with different values for each model were necessary. The range was set to between six and twenty topics with a two-topic interval, as topical divergence was minimal in our dataset.

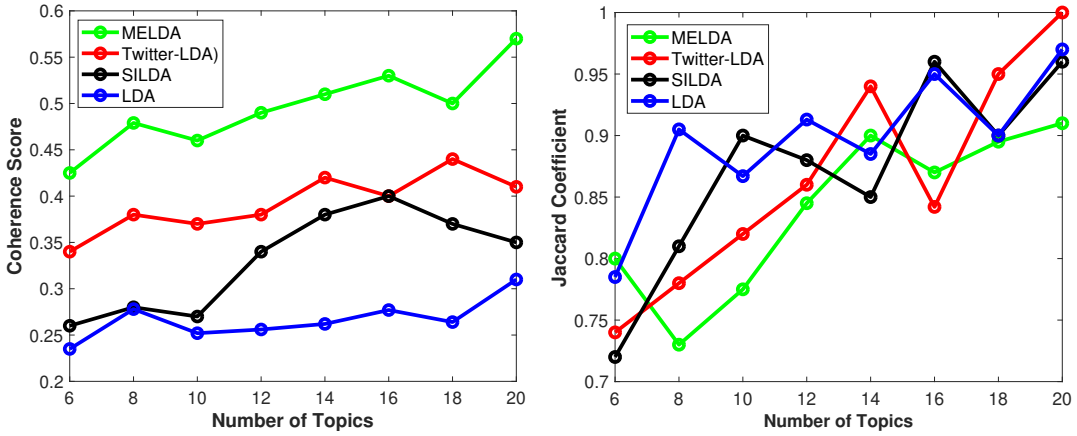
3.4.2 Quantitative Evaluation of Topics

The resultant model was evaluated based on some objective metrics. In studies related to topic modelling, evaluation of the final model involves several techniques. In most topic modelling experimental setups (Blei et al., 2003b), *perplexity* was the measurement metric to evaluate how well the model fits data. However, perplexity was limited in terms of its indifference to human judgment (Chang et al., 2009) as well as its reflection of semantic coherence on individual topics. The goal here was the generation of coherent topics from short texts with augmentation from longer texts.

Therefore, topic coherence (Mimno, Wallach, Talley, Leenders & McCallum, 2011) was applied as the main topic evaluation metric. Human interpretability of extracted topics corresponded well with topic coherence and thus influenced the choice of this measurement metric. Topic distinctiveness was another important metric in the evaluation of modelled topics Therefore, the *Jaccard Coefficient* measure to evaluate the distinctiveness of the most probable words that co-occur in each topic was opted for.

Topic Coherence

Several works in the literature made use of topic coherence in the evaluation of topics (He, Wang & Jiang, 2017b, 2017c; Z. Chen et al., 2013c, 2013a). Topic coherence relied upon word co-occurrence patterns in the corpus. Therefore, it corresponded well to human judgment and in this case interpretability of topics (Mimno et al., 2011). Twenty most probable words in each topic were used and average values for all the topics were computed. A higher topic coherence value indicated better topic quality in relation to



(a) Topic Coherence (b) Jaccard Coefficients

Figure 3.4: Quantitative evaluation results. (a) shows average topic coherence scores for each model. (b) shows the average Jaccard Coefficients for each model.

human interpretability.

The topic coherence model in this instance was a pipeline encompassing segmentation of topics as a representation of the entire set. This measured how single words, or their subsets, fitted together. These were tuples from the top N words in each generated topic. Segmentation was the first step where word-sets were divided into smaller pieces for co-occurrence pattern matching. The second measure involved a probabilistic distribution of the top N words over the reference corpus, which was basically the remaining set. This was at word level where word-pairs W_s were compared against each other. For better coherence computation, there needed to be a confirmation measure indicating an agreement in the word-pairs. Normalised Point-wise Mutual Information (Bouma, 2009) of words denoted as M, was followed. Confirmation measures used word probabilities, which was the third dimension in the configuration space. Lastly, methods related to aggregation of scalar values were computed by the confirmation measures. This was the fourth dimension in the configuration space denoted as α. All values were then aggregated to a single coherence value c. In summary, the computed coherence measure was a pipeline that was a cross product of the four sets $C = (N \times M \times P \times \alpha)$.

Jaccard Coefficient

The Jaccard Coefficient is a widely used measure for quantifying similarity between two finite sets. In topic modelling, it is useful in measuring the overlap of most probable words in generated topics. A lower Jaccard Coefficient value shows better distinctiveness between topics.

Results in Figure 3.4 (a) and Figure 3.4 (b) respectively present the coherence and topic similarity values for each model with topic numbers set between six and twenty with a two-topic interval size. The low value of the topic interval parameter was used to capture small variances between the modelled topics.

The following observations were made :-

1. Topic coherence results presented in Figure 3.4 (a) confirm that MELDA consistently outperformed other baselines with an increase in the number of topics. MELDA performed better than Twitter-LDA as we introduced seed topic sets with a seed confidence value during initialisation. As mentioned in Section 3.4.1, topical diversity was limited because of the dataset nature. Therefore, coherence worsened with the introduction of a larger topic gap in the modelling process. This was the reason for maintaining a two-topic interval size, which was ideal after several trials.

   A higher topic coherence score indicates better model quality, as the generated topics and word co-occurrence patterns conform better to human interpretability. SILDA, LDA and Twitter-LDA coherence scores were consistently lower than MELDA's. For LDA, this could be attributed to its inability to handle noise in tweets, and sparsity in the word co-occurrence patterns. SILDA on the other hand was affected by noisy interest sets and thus performed sub-optimally. Twitter-LDA's performance was somewhat worse when compared to MELDA's. This could be attributed to the noise, as there was no dedicated method to mitigate the

same. Twitter-LDA also lacked guidance in the topic initialisation step, hence the lower scores.

2. A lower Jaccard Coefficient measure indicated better distinctiveness in the generated topics as mentioned in Section 3.4.2. In computing the coefficient, differences between two topic sets were computed for each model at each topic interval. The values were averaged at each interval, which generated the y-axis values in Figure 3.4 (b). MELDA generated the lowest similarity measure, meaning it generated highly distinctive topics. Twitter-LDA reported slightly higher coefficients on average compared to MELDA. However, Twitter-LDA reported an inconsistent result between the 14th and 16th topic model intervals where a 0.1 drop in the Jaccard Coefficient value was noted. Other than that, the rest of the results were consistent with the addition of more topics. LDA and SILDA results were consistent insomuch as they recorded a higher similarity measure, meaning the topics they generated were not very distinctive. It was presumed that the reasons for this performance were more like those in the coherence measure, i.e., noisiness of the dataset and an equally noisy interest set.

3. Generally, more topics negatively influenced the coherence and Jaccard Coefficient values, and thus limited the number of topics to 20. Vocabulary sparsity in tweets as well as low topical variance in the data imposed an upper limit on the number of topics that could be modelled from the data. The fact that the model identified more coherent topics in such a constrained vocabulary space meant that it was effective in modelling more interpretable topics.

3.4.3 Human Evaluation

The primary goal of a topic model is to generate explicable topics that largely conform to human knowledge. Despite the importance of quantitative evaluation, it is not complete

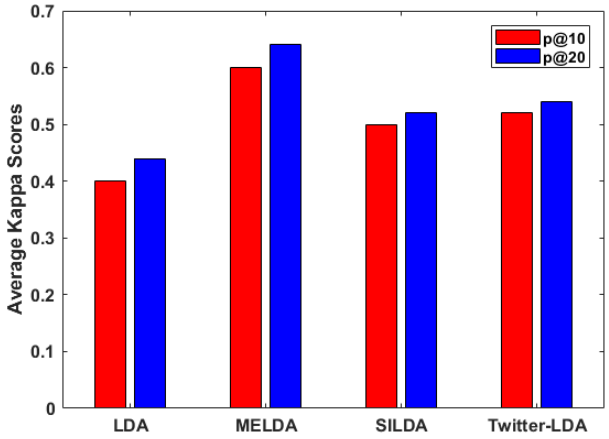


Figure 3.5: Percentage of good topics generated by each model with number of topics words set at 10 and 20 respectively

when judging model performance. Another vital aspect is a qualitative evaluation that is carried out by humans to discern quality of the generated results. To facilitate this qualitative evaluation, five judges with fluency in English and good knowledge of the dataset domain were selected to label every topic generated by MELDA, LDA, Twitter-LDA and SILDA. It is worth noting that the judges are unrelated to the authors and were thus not considered biased. The top 20 words were chosen and ranked by per-topic word distributions for each topic. A word in a topic was considered good only if all the judges agreed to it; otherwise it was labelled as bad. Judges also labelled words that had no clear positive or negative inclination as neutral. For transparency, topics were shuffled so that the judges had no idea which model generated the topics they labelled. In addition, the top-N words in a topic were labelled *good* if they coherently related to the semantic concept represented by the topic. The same approach was used in (He et al., 2017a).

Table 3.1: Cohen's kappa score

	p@10	**p@20**
Kappa	**0.6**	**0.64**

Table 3.2: Sample Topics from Three Models (Good words are in bold and red, neutral ones in green and bad ones in black). Only good and bad word classifications were considered in computing the Kappa Score in Figure 3.5 and Table 3.1

Battery			Screen			Camera		
MELDA	**Twitter-LDA**	**SILDA**	**MELDA**	**Twitter-LDA**	**SILDA**	**MELDA**	**Twitter-LDA**	**SILDA**
battery	**battery**	**battery**	**screen**	**screen**	**screen**	**camera**	**battery**	**screenshot**
note	**note**	**note**	**lock**	**black**	**phone**	**leicacamera**	**camera**	**camera**
samsung	**galaxy**	**samsung**	**ios**	**phone**	**iphone**	**amazing**	**screen**	**screen**
galaxy	**catching**	**galaxy**	**home**	**blank**	**lock**	**capture**	**phone**	**focus**
catch	**causing**	**fire**	**screenshot**	wont	**ios**	**dual**	**screenshot**	**issue**
fire	forbes	iphone	**home**	restart	**black**	**selfie**	**issue**	op
recall	ios	problem	**button**	back	wont	**shot**	oneplus	oneplus
causing	problems	new	app	reset	app	**issue**	one	battery
news	blame	next	music	turn	help	**perfect**	please	xperia
problems	testing	com	**phone**	white	update	oneplus	op	**phone**
Good (80%)	**Good (50%)**	**Good(50%)**	**Good(80%)**	**Good(40%)**	**Good(60%)**	**Good(90%)**	**Good(60%)**	**Good(60%)**
Neutral (20%)	**Neutral (20%)**	**Neutral(30%)**	**Neutral(20%)**	**Neutral(40%)**	**Neutral(20%)**	**Neutral(10%)**	**Neutral(10%)**	**Neutral(30%)**
Bad(0%)	**Bad (30%)**	**Bad(20%)**	**Bad(0%)**	**Bad(20%)**	**Bad(20%)**	**Bad(0%)**	**Bad(30%)**	**Bad(10%)**

Topic Labelling

Judges were tasked to label each word in each topic. Since the judges did not have prior knowledge of the datasets, tweets and related metadata were presented for skimming. This was to help them understand word co-occurrence patterns in both sets. A word was therefore labelled as *good* if it is semantically related to an overall topic concept, otherwise it was labelled as *bad*. *Neutral* words did not fall in the two categories. For simplicity, just the two labels were chosen in for placement into two widely separated categories. The approach by He et al., (He et al., 2017a) of using two sets of probable words, i.e., 20 and 10 word sets to reduce their ambiguity was adopted. Noisy short texts made up most of the top ten most probable words and thus judges found it hard to select the correct concepts. After labelling of words, topics containing more than 60% good words were labelled as *good topics*. To represent this metric, we used Precision@n ($p@n$), where p was the percentage of good topics and n the sum count of good words for each topic word-set, i.e., (n = 10 and n = 20). Results of the above process were represented as the Cohen's Kappa score where all topics with good words above 60 percent were deemed good topics. Table 3.1 depicts MELDA's average scores with different values of n. Figure 3.5 exemplifies the human labelling results from which we can derive the following conclusions:-

1. MELDA performed better than the other baselines on the dataset. Human judges' results showed that MELDA was the only one that surpassed the 0.6 Kappa score implying that it generated *good* topics and related words. Twitter-LDA and SILDA had slight variations in the number of good topics at p@10 and p@20. LDA's good topics and related terms were lower than SILDA's and Twitter-LDA's. The noisy interest preference set affected SILDA's performance as the resultant topics depicted redundancy, as pointed out by the judges. Lack of external knowledge in Twitter-LDA and LDA also affected their performance.

2. Judges noted that an increase in the number of words for evaluation increased their overall conceptual understanding of the topics. Tweets were noisy and the top-10 words in each topic did not exhibit good coherency, in general. Judges noted more descriptive terms with addition of words in all models. This explains the reason for better judgment with $(n = 20)$

3. Topical variance was an issue that the judges noted. It was especially constrained by the reuse of words in different contexts that express different aspects of interest. For example, a word like *replacement* was predominantly used in relation to *battery* and *screen* topics. This was attributed to the dataset nature but a quick look through the external data enhanced their understandability of different co-occurring terms in different aspects.

   Table 3.2 shows sample topics from three models. The top-10 ranked words from MELDA, SILDA and Twitter-LDA are listed. LDA performance was low and thus its top-10 word list was omitted in the table due to space limitations. MELDA modelled topics better compared to the other approaches. We specifically chose topics related to *battery*, *screen* and *camera* issues in the dataset. MELDA performed better in the identification of words coherent to the topics. Words like *selfie* and *shot* in *camera* topic were only captured by MELDA. The same result

was replicated in the *battery* and *screen* topics. Twitter-LDA had a slightly higher number of good words compared to SILDA. Experimental results from human judges demonstrated MELDA's good performance in the extraction of coherent topics and related topical word distributions in a topically constrained corpus.

3.5 Model Generalisation Limitations and Workarounds

The model as described in Sections 3.1 and 3.4.1 is of the assumption that the metamodel is formulated from a known external data source. However, this approach also has the following limitations.

1. **Choice and availability of external data** - As described in Section 3.4.1, availability of external datasets is domain specific. Some domains have readily available external data. For example, smartphone or movie related tweets can be correlated with Amazon or IMDB reviews while others do not have, e.g., emerging events such as presidential elections, terror attacks or even earthquakes. This presents challenges in applying this modelling process. However, there are two possible solutions to this issue:-

   (a) Amalgamation of external data sources - Depending on the nature of the short texts, several external dataset sources can be combined. In the case of Amazon reviews as in Section 3.4.1, several review sources can be amalgamated to enhance the semantic relevance of the external set. For example, reviews from other websites [7],[8] combined with Amazon ones are ideal as long as the entity of interest is the same. Entities such as *battery life*, *camera*, *screen size* and related information are likely to have the same semantic relevance across the external sources.

[7] `https://www.bestbuy.com`
[8] `https://www.gsmarena.com/`

(b) Some short-text datasets do not have semantically relevant external sources. In this context, a metamodel can be built by merging topic vectors from generalised and task-specific KBs such as Wikipedia, DBPedia, etc.

2. **Linguistic and Semantic Equivalence in the datasets** - The semantic and linguistic distance between the datasets is imperative. Conventionally, MELDA, just like LDA, automatically identifies topics based on term co-occurrence likelihood. This presents a challenge in convergence towards the seed topics in the event the datasets are linguistically diverse. This may be in terms of language diversity or low term co-occurrence likelihood. In this case, **self-learning**, where short texts can be enhanced, learned and modelled, is a viable solution. Deep learning related methods to some extent have proved to work in such scenarios, especially after some textual enhancements (Zhan & Dahal, 2017).

3.6 MELDA's Significance

MELDA's potential use is in several application scenarios: -

1. **Content Recommendation** - Third party content recommendation is one such potential application domain. Content creators may be interested in contacting users whose tweets are inclined towards certain seed documents that they created. Such documents in this instance formulate the metamodel whose seed topics can be used in modelling better topics from the tweets.

2. **Building of ontologies** - Incremental learning is pertinent in the construction of, for example, term-based ontologies, where documents and related terms are mapped in the same space. This is vital in understanding other related documents which, in this case, can be the short-text documents. Such tasks are related to topic modeling as well as understanding document semantics.

3. **Automatic term-based tagging** - Content creators would be interested in users with tagged content that is relevant to them. In such a situation, the tagging process is modelled better when provided with documents to base the tagging process on.

4. **Cold-start scenarios** - Gathering enough content before making recommendations is hard, especially in user-specific microblog data, such as an insufficient number of tweets for a certain user or group of users. Modelling related external data can help in the provision of content for the initial recommendations.

3.7 Chapter Summary

MELDA, a novel semi-supervised approach for modelling of coherent topics and respective word distributions in short texts was proposed in this section of the thesis. It is based on LDA that is probabilistic. By incorporating metamodel seed topic sets and the seed confidence value at the initialisation step, MELDA was able to steer short-text words and their co-occurrence patterns towards the metamodel ones. A metamodel was formulated from a long-text dataset, contextually related to the short-text dataset. In MELDA's generative process, short texts in the form of tweets were distributed over metadata topic labels generated via LDA. A seed confidence value biased this distribution at the topic initialisation step based on word co-occurrence patterns in the metamodel topics. It was demonstrated in the experimental results that the approach quantitatively and qualitatively outperformed three other state-of-the-art approaches in terms of topic coherence and distinctiveness. Human judgment results also proved that MELDA generated more coherent topics. Notable limitations in this research are in relation to the lack of a general procedure for acquisition of prior knowledge, as well as linguistic and semantic relevance between the two sets of data.

Chapter 4

Follow-Back Recommendations

4.1 Introduction

This chapter extends the short texts knowledge extraction phase detailed in Chapter 3. The assumption is that once a better methodology to extract knowledge in short texts exists, it becomes easier to extract more representative user interests. In essence, the user interests extraction problem is approached by modelling follow-back recommendations that ideally are interest-dependent. Follow-backs represent the bi/uni-directional relationship between users on short text microblogs. Normally this followership is determined by the convergence of interests between the users. Is it possible to recommend follow-backs based on the interests of other users on short-text microblogs?

For example, users on Twitter tend to consume certain content to a greater or less extent depending on their interests over time. Quantifying this degree of content consumption in certain topics is an arduous task. This is further compounded by the amount of digital information that such platforms generate at any given time. Formulation of personalised user profiles based on user interests over time and friendship network is thus a problem. Therefore, user profiling based on their interests is important for personalised third-party content recommendations on the platform. To put this in

perspective, a user may be interested in *political content*. To recommend a friendship network to this user, the recommendation engine has to *i) decipher that the user has politics as his/her interest ii) the friendship connection (follow-back) to be recommended has to be in the same user interests semantic space*, i.e., needs to be interested in politics at least to a certain threshold.

This problem is addressed in this chapter as a two-step process:- *(i) Firstly, users' Degree of Interest (DoI) towards a certain topic based on the overall users' affinity towards that topic is computed. (ii) Secondly, this DoI is affirmed by correlating it to their friendship network*. Furthermore, the model for DoI computation and follow-back recommendation system is formulated by learning low-dimensional vector representation of users and their disseminated content. This representation is used to train models for the prediction of correct topical cluster classifications. In the experiments, a Twitter dataset just as in Chapter 3 is used to validate the approach by computing degrees of interest for certain test users in three diverse and generic topics. Experimental results showed the effectiveness of the approach in the extraction of intra-user interests and better accuracy in follow-back recommendations with diversities in the topics.

4.1.1 Notations

The below notations and respective descriptions are used in this chapter: -

D_w - Dictionary D of n-grams with words w.

Z_d - Vector representation associated with each n-gram d.

c - Context position of a word in the vocabulary.

W_x - Set of words in a tweet x.

w_x - Vector representation of a word in word set W_x.

$w_x^{'}$ - Model for vector representation w_x.

Y - Set of vector representations of centroids (centroid map) for clusters of interest .

T_u - Timeline for tweeter u.

x_u - Extracted tweets from T_u.

$DoiSCC$ - A user's Degree of Interest in certain topics.

4.2 Problem Statement

User interactions and proliferation of citizen journalism on microblogs such as Twitter and Facebook has led to the generation of enormous amounts of online content. The disseminated content on these platforms is also diverse, e.g., text, videos, images or intra-user interactions via 'retweets', 'mentions' or 'likes'. Considering the nature of the content and dynamism in intra-user connections, and specifically on Twitter, it is difficult to infer users' interests. Ability to solve this issue is important in recommender systems research.

Typically, recommender systems augment this personalization process by suggesting meaningful follower-followee relationships. On Twitter, user interests are largely extrinsic, based on their declared interests. However, dynamism in the disseminated content as well as changes in follower-followee relationships imply that user interests change over time. This poses the following questions : -

1. Is it possible to profile a user/group of users based on their disseminated microblog posts over time?

2. Does homophily in microblog platforms friendship networks influence formulation of online user profiles?

Furthermore, interests that microblog users have towards certain topics change over time, e.g., hashtags based on prevailing topics at the time. Regarding user's friendship networks, it is not correct to assume that a user and his/her followees share the same interests albeit to a certain threshold. It is typical for example for some microblog users

to have many followers but not all of whom are followees. However, there is a need for such users to create a network of influencers to propagate their content as well as to receive relevant content from other actors with the same interests.

Therefore, a framework is presented to compute the *Degree of Interest* that users may have towards a certain topic based on their disseminated content over time. As a case study, interest in *sports betting*, *Swahili related chatter* and *daily news chatter* are considered as topics of interest in a generic Twitter dataset. The framework quantifies the degree of interest in a specific microblog topic by analysing user's short, sparse and noisy content over time. More precisely, this chapter's contributions are as below :-

1. A framework is developed that formulates user profiles based on user-disseminated short and noisy texts.

2. Topical affinity computation in short texts. Affinity for the topics of interest is what ultimately determines the clustering of users in the same embedding space. In this respect, the modelling approach can identify user-representative interests, more so in short and often misspelled words as evidenced in Section 4.3.1. This presents quality clusters whose centroids are pivotal in computation of affinities from test users in Table 4.2. This computation process is different from the works in user interests and preferences modelling as detailed in Section 2.4.3

3. Follow-back recommendations. The process of determining the most representative users to follow-back based on a user's distance to the centroids of interest distinguishes this work from research in the topical recommendations domain in Section 2.4.4. This is complemented by the application of the *Theory of homophily* as in Section 4.4.3. It was possible to accurately identify contextually representative users as proven in Section 4.5.3 by real tweeters.

4. The framework is validated across three high level topics of interest, i.e., *sports*

betting, *swahili related chatter* and *daily news chatter*. An extensive experimental study in formulation of user representative profiles is done by analysing the disseminated content over time.

Access to Twitter related content in specific topical classifications contributes to a better understanding of tweeter's social patterns and content. The goal in this research aspect is with regard to improvement of application methods in the evaluation of the degree of interest that users may have towards a topic. Once computed, this metric can be used as input in modelling the tweeter's interest in topics such as sports betting.

Low-dimensional vector representations in form of sentence embeddings are utilised to comprehend the underlying semantic structure of tweets to profile tweeters more accurately. In addition, a responsibility matrix representing individual interests in topical clusters as a cosine similarity measure in the computation of the interest levels in certain topical content is generated. Neural network representations in this instance worked well with misspelled/shortened words. This is a common occurrence in tweets as well as making intelligent guesses for out-of-the-vocabulary words that had some form of character-level consistency with the terms in the vocabulary.

4.3 Follow-Back Formulation and Summary of Literature

The core processes for the framework to automatically compute the degree of interest in microblog topics of user interest are presented. The framework encompasses computation processes related to *short-text modelling* as well as *follow-back recommendations* [1]. Comprehension of short and informal texts reminiscent of microblog posts presents a challenge for modelling algorithms (S.-H. Yang et al., 2014b), (W. X. Zhao, Jiang,

[1] This work has been published in the Information Systems Journal, and presented at the 2020 Hawaii International Conference on System Sciences

Weng et al., 2011). This is despite their success in modelling of conventional texts (Blei et al., 2003a). A neural network approach was adopted in modelling text, more so at character level due to the nature of the short texts. Their success in modelling such texts has been demonstrated in the literature (J. Li, Xu, He, Deng & Sun, 2016), (Mishra, Rizoiu & Xie, 2018), (Zhang, Robinson & Tepper, 2018). An in-depth review of short-text modelling and follow-back formulation literature is in Section 2.4.4.

Generally, most short-text microblog posts are made up of shortened and misspelled words, in addition to the informality in the language of expression. Therefore, the following processes were influenced by the nature of the dataset and research problem.

1. **Short-text modelling** - A corpus of tweets is used to train a model using *FastText*[2] algorithm. The model generates low-dimensional vector representations of words in the corpus (Bojanowski, Grave, Joulin & Mikolov, 2017). In *FastText*, the model learns embeddings at character-level thus relevant to the nature of our dataset (Bojanowski, Grave, Joulin & Mikolov, 2016). This modelling approach is of choice after a performance comparison with other state-of-the-art methodologies as shown in Table 4.2 in this chapter.

2. **Extraction of centroids and clustering** - A clustering approach is used to group similar tweets by averaging vectorised intra-word similarities. In this case, cluster centroids are used to infer clusters for new tweets depending on the semantic distance between the model's and test tweets vectors.

3. **Computation of users Degree of Interest (DoI)** - Sample tweets are vectorised via the trained model and the proposed method is applied to measure tweeter's level of interest in the topic based on the disseminated tweets over time.

4. **Association of tweeter's DoIs with their friendship networks** - A correlation

[2] `https://fasttext.cc/`

measure of a selected list of tweeters and their friendship network is computed. This follows the *homophily* theory where users with similar interests tend to relate, an important process in modelling follow-back recommender system.

4.3.1 Short-Text modelling

The adapted modelling methodology was primarily based on the neural language model, *FastText* (Bojanowski et al., 2016). The model works by extracting syntactic information in a textual corpus independent of the language of expression, vocabulary size and misspelling reminiscent of tweets. *FastText* creates a model with a vector representation of a word(s) based on the context within which the word is commonly used. To validate the modelling approach, the same dataset was trained with *Word2Vec* (Mikolov, Sutskever, Chen, Corrado & Dean, 2013) and *Glove* embeddings (Pennington, Socher & Manning, 2014) baselines with different but consistent dimensions across the frameworks. *FastText*, unlike other word embedding algorithms, does not ignore the morphology of words. This is a limitation especially in languages with a large vocabulary as well as those ones with rare words. With *FastText* modelling, a vector representation is associated with each character n-gram, where words are modelled and represented as the sum of character vectors in those words. This makes it the most appropriate algorithm in dealing with mispelled and out-of-the-vocabulary words. It also uses a sliding window of characters in computing word vectors.

Vector words representations make it possible to infer their contextual usage. This is important in computation of inter-word/sentence distances. Words that are contextually similar are usually used together in expressions like *"good morning"*. Semantically, such words will have a higher similarity score, i.e., closer to 1 than compared to dissimilar ones. Therefore, a cosine value close to 0 depicts very low term/sentence contextual similarity. We followed the following process to model tweets :-

- **Pre-processing** - For better analysis of unstructured short texts, preprocessing is necessary. For instance, prepositions and punctuation in documents do not provide any meaning, more so contextually thus were filtered from the dataset. To clean up the tweets which form the model training corpus, the following steps were followed :-
  - Lower-cased all words for uniformity.
  - Removed all numbers and encoding accented characters.
  - Removed all hyperlinks in the corpus.
  - Removed all hashtags, as they were not of interest in this instance. Hashtags are just words inserted manually by tweeters in a tweet and are prefixed by the hash (#) symbol. Their function is to help identify topically similar tweets.
  - Removal of user mentions. Syntactically, they are presented just like hashtags except that they are prefixed by the @ symbol.
  - Removal of words whose length was less than three characters. Their contextual relevance was not significant in successive experiments.
  - Removal of stopwords. We used the NLTK stopword list [3].
  - Tweets tokenisation, where individual terms in each tweet are split and appended in a list for modelling.
- **Model Training** - A cleaned and tokenised corpus of tweets was the input in the model training pipeline across the state-of-the-art frameworks mentioned above. The models were trained to learn conceptual knowledge of the dataset by mapping each word to a continuous vector space from its distributional properties observed in the corpus.

[3] http://www.nltk.org/

The following parameters were specified in order to train *FastText* and *Word2Vec* models:

- *size* or the number of dimensions;
- *min_count* or minimum count of a word in the corpus for it to be factored in the training;
- *sg* for training a skip-gram model if sg = 1, otherwise Continuous Bag of Words (CBOW).
- The *window* parameter specified the maximum distance between the current and predicted word in a tweet;
- *word_ngrams* to enrich word vectors with subword(n-grams) information if specified as 1 ; and *iter* or iterations which was the number of iterations (epochs) over the corpus.
- *Glove* model only provisions for the *epochs* and *learning rate (lr)* to be defined. The rest of the parameters were default to the model.

The model outputs were vectors for each word in the corpus. Since vector representation in *FastText* is linear, additive compositionality was possible in *FastText* based models. In addition, *FastText-CBOW* uses a distinct vector representation for each word, whereas the skip-gram model ignores the internal structure of words. Therefore, each word w is represented as a bag-of-character *n-grams*. The representation in FastText is such that the word itself is also included in the set of n-grams. *Word2Vec* vector representation works the same way, albeit at word-level. In the implementation, $3 \leq n \leq 6$ window of characters was implemented in the FastText based frameworks. This way, most of the n-grams were considered, based on the average word lengths in tweets. Given a dictionary of n-grams, i.e., size D and a word w, let $D_w \subset \{1, ..., D\}$. A vector representation

z_d is associated with each n-gram d. Therefore, a word is represented as the sum of n-gram vectors. The scoring function as in (Bojanowski et al., 2017) is formulated as below in this summation process :-

$$s(w, c) = \sum_{d \subseteq D_w} z_d^{\top} v_c \tag{4.1}$$

where c is the context position of a word, and v the corresponding word vector.

A pre-processed tweet ready for vectorisation is made up of word *tokens*, i.e., a list of all words split up per tweet. Therefore, its representation is the sum of the tweet's word vectors, as they are linear. In this case, let W_x be the set of words in tweet x. The words are then vectorised. Therefore, w_x is the vector representation for a given word in the set. The tweet model $w_x^{'}$ is then represented as below:-

$$w_x^{'} = \sum_{w_x \subseteq W_x} w_x \tag{4.2}$$

Specific values for training the model are elicited in Section 4.3.1. At the end of this process, the model is ready for extraction of vector representations for each word in the corpus.

4.3.2 Extraction of Centroids and Tweets Clustering

A correlation of test tweets to the clusters of interest is computed in this step. To delimit the cluster of interest vector space, words in the cluster have to be grouped based on a semantic distance measure. Generally, clusters are defined via approximation and manual inspection of the underlying keywords. Each cluster is represented as a semantic topic where its keywords define it. A tweet representation $w_x^{'}$ is used to define cluster numbers. In the event clusters were to be labelled, then manual inspection of defining keywords in each cluster was used to define cluster names.

To cluster tweets, *K-Means++* (Arthur & Vassilvitskii, 2007) was applied on the training corpus. This clustering algorithm optimises the choice of cluster centres for k-means by spreading out the initial set of cluster centroids so that they are not close to each other. With the initialisation of k-means++, an $O(\log k)$ solution was guaranteed. On the other hand, FastText uses a hierarchial softmax to reduce the computational complexity in the profiling process to $O(hlog(k))$. k is the number of classes and h the dimension of text representation. This made it possible to find the best set of centroids. The clustering insight is that objects within the cluster are as similar as possible, whereas those from different clusters are as dissimilar as possible. An optimal clustering process in most cases is dependent on the final purpose of the clusters, i.e., the level of detail required of the clusters. To determine the optimal cluster numbers, a heterogeneity convergence metric on the models (Bholowalia & Kumar, 2014) was utilised. To determine the metric, tests were run considering different k values (cluster numbers) on the dataset. The number of clusters that best identified the dataset at the elbow point were selected. The cosine distance measure was then used to compute the intra-cluster distance between y points in each cluster Z_k and centroid Z_z in that cluster.

The main aim of this phase was to identify words normally used in the cluster of interest and eventually compute its centroid map. Computation of centroids, albeit by word as in (Recalde & Kaskina, 2017), provided a reference point for computation of the distance between clusters and any other tweet.

4.3.3 Computation of a User's Degree of Interest (DoI)

To understand how the degree of interest in a cluster was computed, similarity of test tweets with respect to cluster centroids was first computed.

- **Tweet Similarity to Cluster Centroids:** Computation of cluster centroids was important in classifying tweets with respect to the clusters. Let Y be the set

of vectors representing centroids for clusters of interest, i.e., $t \in Y$. Y was important in measuring the similarity of any given tweet to the generated clusters. In this case, the interest was to find the semantic distance between a tweet and clusters that semantically represented *sports betting, Swahili chatter* and *daily news chatter* clusters t_{SSC}. Analogy tests and manual inspection of keywords in each cluster as in Section 4.4.1 informed the topical names of these clusters. Choice of cluster numbers was informed by the results in Table 4.2. Given a tweet x, let W_x be the set of words in the tweet. To find the vector of x, the sum of vectors for words in W_x as represented in Equation 4.2 was computed. Cosine similarity was applied to measure the semantic similarity between a tweet x, represented as vector $w_x^{'}$ to cluster centroids $t \in Y$ as earlier defined (Saxena et al., 2017). Cosine similarity was relevant in this case as similarity between tweets and the cluster centroid maps was possible irrespective of their vector sizes. This is because it measured the cosine of the angle between the two vectors ($w_x^{'}$ and $t \in Y$) projected in a multi-dimensional space. The advantage with cosine similarity measure is that despite the two vectors sets being far apart by say Euclidean distance due to their size, chances are that they may still be semantically close. The smaller the angle between the two sets, the higher the cosine similarity. Therefore, two objects are presumed very similar if the cosine distance is close or equal to 1 and dissimilar if close or equal to 0.

$$s_{xt} = CosineDistance(w_x^{'}, t), \forall t \in Y \tag{4.3}$$

Equation 4.3 presents the cosine similarity computation. Therefore, it was possible to compute the semantic relevance of a tweet to clusters ($s_x t_{SSC}$).

- **DoI Computation:** Computing the DoI of a given tweeter u involved a few steps. *(i)* Tweets that user u posted on his/her timeline T were extracted via Twitter's

Search API[4]. The API has the capability of returning a maximum of 3200 tweets for any given Twitter account. *(ii)* The second step involved pre-processing of tweets $x_u \in T_u$ as in Section 4.3.1. The similarity of the extracted tweets x_u to the clusters of interest t_{SCC} was then computed. The Degree of Interest in the Sports betting, Swahili and Daily news related content clusters (DoiSCC) was calculated as the average of vector similarities $s_{x_u t_{SCC}}$. Therefore,

$$DoiSCC_u = average(s_{x_u t_{SCC}}), \forall x_u \in T_u, t_{SCC} \in Y \tag{4.4}$$

Since the cosine distance was used, therefore, a tweeter's DoiSCC value close to 1 in relation to any of the clusters meant that the user's tweets were mostly inclined towards that specific topic. On the other hand, values close or equal to 0 meant minimal user interest to the domain. Algorithm 4 details the above processes.

4.3.4 Tweeter's DoiSCC vs their Friendship Network

A tweeter's profile is defined by the content he/she shares as well as with the intra-user relationships that exist from the 'mentions', 're-tweets' and 'follower' relationships.

For evaluation purposes, the degree of interest for the friendship network of the initial users was computed. This followed same process as in Algorithm 4 after collecting tweets from the friendship network timelines. Friendship network in this case was any tweeter who was "mentioned","retweeted","liked" or "followed" the original tweeters and related content. These are the tweeters who disseminated the 298835 tweets as in Section 4.4.1. To obtain tweets from the friendship network, these processes were followed:-

1. Searched for tweeters fulfilling the friendship network conditions as above.

[4] `https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.html`

2. Downloaded a maximum of 3200 tweets and related metadata per user using Twitter's search API [5].

The friendship network's DoI computation process follows the same modelling pattern as in Section 4.3.3. The end result is a computation of the friendship network DoIs and a comparative evaluation with the original user's DoI as in Section 4.5.1. This informs the DoI recommendations thresholds for users and their friendship networks. The results are fundamental in the design of short-text based recommender systems for users, third-party content or groups and lists.

In Algorithm 4, inputs in the modelling process are tweet tokens W_x, that are vectorised and summed as $w_x^{'}$ (Line 2 - 4). Initialisation to find cluster centers for word vectors W_x is the first step in the clustering process (Line 1). Similarity of a tweet to a centroid t is computed by measuring the cosine angle between tweet vectors and the centroid of interest in centroid map Y (Line 8). The centroid map is a list of terms and respective cluster labels. The cosine distance is computed for all tweets under each user (Line 12). To compute the DoI, a summation of individual tweet DoIs for each user is made. The average DoI (DoiSCC) across the tweets, i.e., $DoiSCC = \{\sum DoI\}/len(x_u)$ (Line 13) is the result in the computation, i.e., a representation of the user's as well as friend's interest in the topic of interest.

Computation of the friendship network $DoiSCCs$ followed the same process as in Algorithm 4 except that the input was tweeters' usernames in their respective clusters as illustrated in Section 4.3.4.

This process is illustrated in the computational workflow in Figure 4.1 with four sub-sections. Tweeters consisted of the initial tweeters' usernames, with their corresponding collection of tweets. The collection was limited to 50 tweets per user due to Twitter search API limitations. Pre-processing of the tweets text segment followed and resulting

[5] https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets

Algorithm 4 User DoI Computation Process

Input: Tweet Tokens W_x, Word vectors w_x, Tweet Model $w_x^{'}$, Cluster of Interest t, Centroid Map Y
Output: Tweeters DoI U_{DoI}

Initialisation via K-Means++ ▷ Word Vectors w_x
Computation of Tweet Model $w_x^{'}$
for $w_x \in W_x$ **do**
    $w_x^{'} = \sum_{w_x \subseteq W_x} w_x$ ▷ Tweet Model as the sum of word vectors w_x in tweet x
end for
Similarity of a tweet to a cluster of interest in Centroid Map
for $t \in Y$ **do**
    $s_{xt} = Cosine(w_x^{'}, t), \forall t \in Y$ ▷ Computation of the cosine distance between the tweet's model $w_x^{'}$ and cluster of interest t in the Centroid Map Y
end for
Degree of Interest (DoI) Computation
for $x_u \in T_u$ **do**
    $DoI+ = 1 - Cosine(x_{uv}, t_{SCC})$ ▷ Similarity via cosine measure between tweets x_u with the clusters of interest t_{SCC}.
    $U_{DoI} = \{\sum DoI\}/len(x_u)$ ▷ Compute the Degree of Interest (DoI) for the tweeter by computing sum of a user's tweet level DOIs DoI divided by the count of user's vectorised tweets x_u.
end for

tokens being the input to the vectorization process under the modelling sub-section. The resultant vectors with the optimal number of clusters were the input to the clustering process under DoiSCC computation. In clustering, centroids were extracted and the cosine distance to the cluster of interest measured representing the topical interest levels. This process is described in Algorithm 4. The friendship network as the input in the figure, consists of all users who were *mentioned by tweeters*, *replied to their tweets, their bidirectional followers* and those who *retweeted their content*. Detailed individual steps are described in Section 4.3.

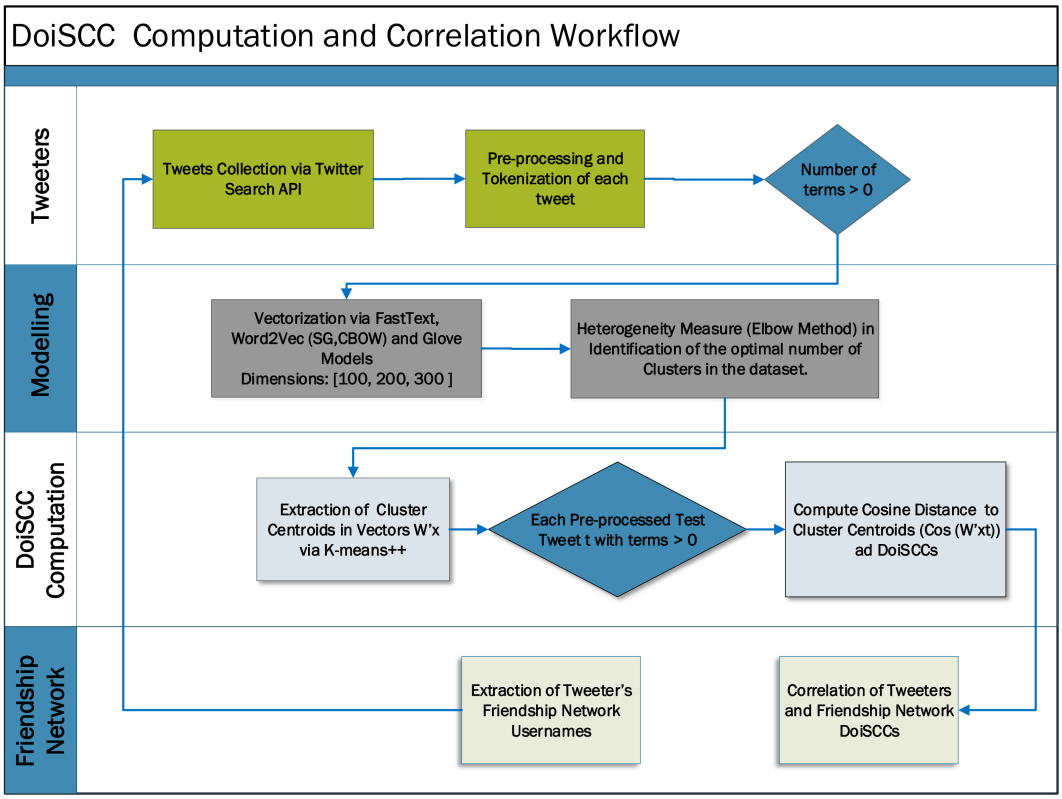


Figure 4.1: Computation Workflow to Generate Tweeters and Friendship Network DoiSCCs

4.4 Experimentation

This section elicits the processes followed to validate the proposed approach presented in Section 4.3. The collected datasets for both tweeters and their friendship network mirrored real Twitter intra-user interactions. The data collection and experimentation processes are described in the sections below: -

4.4.1 Datasets, Settings and Analogy Tests

298835 unique and generic tweets were collected in JSON format for a period of six months starting 1/9/2018. Each entry in the dataset was a single tweet with its associated metadata, i.e., geo-location, mentions, hashtags, etc. Retweets were filtered out from the collection. 90% of the collected tweets were written in English and 6% in Swahili.

The remaining set of tweets comprised of a mixture of English and Swahili vocabulary.

Ground Truth - Sports Betting Data

To accurately measure the topical interest in the generic dataset in Section 4.4.1, ground truth was required. Therefore, the dataset was further curated by adding *sports betting* tweets to the generic dataset that we had collected. The purpose of this process was to make sure that a ground truth cluster existed in the dataset. This was important in the overall evaluation of the modelling framework. Therefore, a pool of 50639 unique sports betting related tweets were added to the generic dataset. These unique betting related tweets were collected from timelines of a few sports betting companies. They included tweets associated with *sportpesa*[6], *betin*[7],*eazibet*[8], *betika*[9] and *betwayke*[10] Twitter handles. The total number of tweets in the training corpus was 349474.

Analogy Tests

Fact checking through *analogy* tests validated the generalisation and quality of the corpus and trained models. Therefore, several qualitative tests were conducted on the optimised models to ascertain their relevance to the test scenario. Table 4.1, summarises validation examples in the context of *sports betting* and *politics* respectively. In these diverse examples, *odds*[11] is a term used in the betting industry. On the other hand, *uhuru*[12] is a Kenyan politician. Terms were used to compute inter-term similarity using the models. Top five most similar terms per cosine distance were then generated per model and with different dimensions as shown in Table 4.1. It's worth noting that the values in this table are simply guidance to the generalisation process for users and

[6] https://www.sportpesa.org/
[7] https://www.betin.co.ke/
[8] https://www.eazibet.co.ke
[9] https://www.betika.com/
[10] https://www.betway.co.ke
[11] https://mybettingsites.co.uk/learn/betting-odds-explained/
[12] https://en.wikipedia.org/wiki/Uhuru_Kenyatta

do not depict comparative inter-model accuracy. For example, some words may have higher values as distance to the test terms though may be irrelevant contextually to the term. Therefore, the analogy tests in Table 4.1 just guided model choices for further evaluation by the authors.

Table 4.1: Sample analogy test with two terms; *odds* - a betting markets related term and *uhuru*, the president and politician in Kenya in models of *100,200* and *300* dimensions.

	Most similar to "odds" - (Betting Term)						**Most similar to "uhuru" - (Politician in Kenya)**					
Model Dimensions	**100**		**200**		**300**		**100**		**200**		**300**	
	Word	***Relevance***	***Word***	***Relevance***	***Word***	***Relevance***	***Word***	***Relevance***	***Word***	***Relevance***	***Word***	***Relevance***
Glove	finns	0.84	finns	0.79	finns	0.76	wheelpic	0.95	wheelpic	0.94	wheelpic	0.92
	available	0.66	acesse	0.61	acesse	0.582	graftheadlines	0.83	opanga	0.77	opanga	0.74
	acesse	0.65	defy	0.60	noen	0.581	opanga	0.80	graftheadlines	0.76	graftheadlines	0.73
	defy	0.64	noen	0.59	defy	0.576	kenyatta	0.79	deactivated	0.75	exits	0.72
	competition	0.62	competition	0.58	competition	0.55	deactivated	0.76	exits	0.73	scolds	0.64
Word2Vec-SG	2odds	0.66	price	0.50	sportpesatips	0.49	uhuru	0.65	ruto	0.60	deactivated	0.44
	guaranteed	0.65	markets	0.443	betnba	0.48	Kenyatta	0.59	Uhuru	0.55	kenyatta	0.439
	bets	0.634	bets	0.44	chekiodds	0.46	kamama	0.56	dp	0.50	inaleta	0.423
	price	0.633	stake	0.43	in-play	0.459	corrupt	0.55	aache	0.46	ocsragira	0.411
	2020qualifiers	0.63	evens	0.40	evens	0.45	president	0.545	murkomen	0.44	puga	0.407
Word2Vec-CBOW	stake	0.58	price	0.50	price	0.46	ruto	0.64	ruto	0.52	ruto	0.55
	price	0.57	markets	0.44	stake	0.42	uhuru	0.63	Dp	0.47	dp	0.45
	games	0.53	bets	0.42	markets	0.40	dp	0.58	murkomen	0.44	alisema	0.409
	bets	0.52	evens	0.41	bets	0.37	ruto	0.54	jubilee	0.436	hamjui	0.408
	markets	0.50	prices	0.39	evens	0.36	raila	0.50	raila	0.428	murkomen	0.40
FastText - CBOW	2odds	0.72	2odds	0.72	2odds	0.70	uhuruto	0.82	uhurus	0.793	uhurus	0.808
	bestodds	0.69	3odds	0.67	3odds	0.68	uhurus	0.81	uhuruto	0.77	uhuruto	0.75
	80/1	0.68	bestodds	0.64	oddset	0.611	uhurukenyatta	0.76	uhurukenyatta	0.72	uhurukenyatta	0.652
	3odds	0.67	odds-on	0.63	odds-on	0.60	kenyatta	0.756	huru	0.68	huru	0.648
	chekiodds	0.66	chekiodds	0.62	chekiodds	0.583	ukenyatta	0.698	kenyatta	0.65	uhuruhighway	0.61
FastText - SkipGram	2odds	0.97	2odds	0.968	2odds	0.966	uhuru	0.88	uhuru	0.89	huru	0.89
	3odds	0.965	3odds	0.967	3odds	0.965	odm-Uhuru	0.87	odm-Uhuru	0.86	odm-Uhuru	0.86
	chekiodds	0.88	chekiodds	0.87	odds-on	0.889	ushuru	0.869	ushuru	0.85	ushuru	0.847
	oddset	0.855	oddset	0.86	chekiodds	0.868	uhurus	0.855	uhurus	0.84	uhurus	0.836
	bestodds	0.76	bestodds	0.74	oddset	0.859	uhuruto	0.85	uhuruto	0.82	uhuruto	0.82

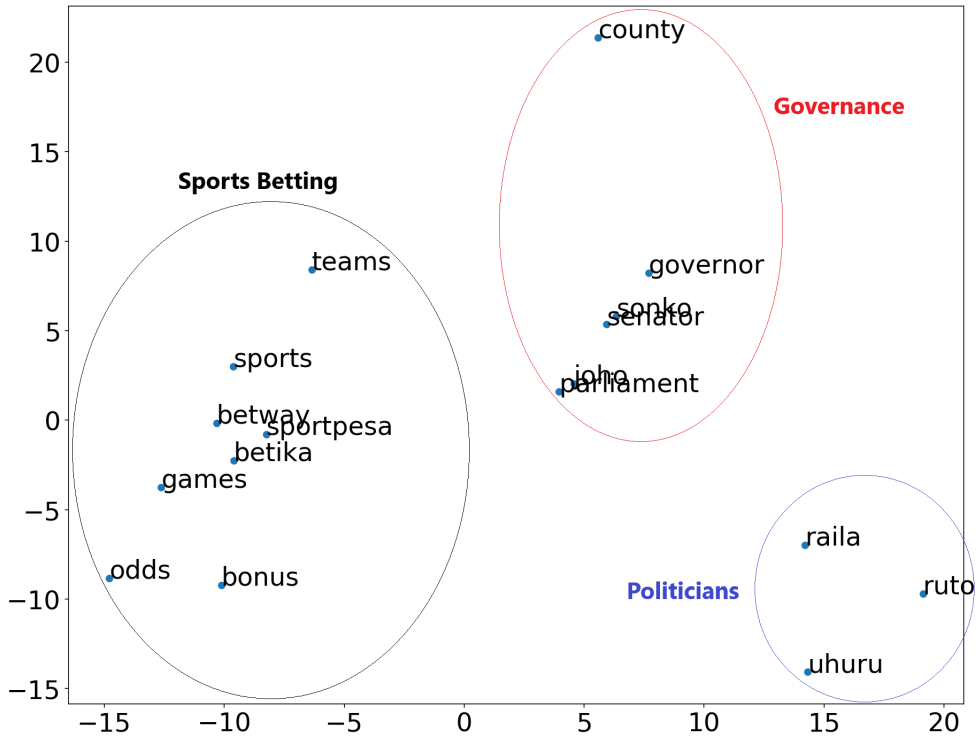


Figure 4.2: Sample plot showing the semantic relevance of words in the training set. Semantic distance between words is depicted by the closeness of the words

Furthermore, FastText-100 model was used to plot the semantic distance between words as part of the analogy tests. Figure 4.2 shows the distance between words in the corpus. Words like *ruto, raila, uhuru* being grouped close to each other is semantically relevant since they are all politicians in Kenya. On the other hand, words like *county,governor,joho, senator, parliament, sonko* are all governance-related. In fact *Sonko* and *Joho* are current governors in Kenya. *Sportpesa,betway,betika* are sports betting companies in Kenya, thus grouped together in addition to words like *odds, bonus,games,sports* and *teams* being semantically close.

4.4.2 Ground Truth Tweet Samples

Analogy tests in Section 4.4.1 provided a general semantic overview of the dataset. However, before selecting the best modelling framework for the research problem, the

models had to be subjected to a known dataset. Therefore, 1000 tweets were sampled from the five Twitter handles of betting companies mentioned in Section 4.4.1. Manual inspection of the 1000 tweets was done by three judges to make sure that all tweets were semantically close to the sports betting domain at least by human understanding. This sample dataset was set up purposefully to test the accuracy of the neural network approaches in terms of best parameter sets in relation to cluster numbers. Cosine distances between the sample tweets and the centroids of the sports betting clusters in models were then computed. This part was significant in achieving following two objectives:-

1. Identification of the most representative model and dimension sizes in training the models.

2. Identification of the number of clusters that depicted the best representation of the corpus.

The assumption in this case was that the higher the number of correctly classified tweets in relation to the sports betting cluster, the better the parameter adjustments. Therefore, it was a matter of iterative trialling of varied parameters in the models. Cluster assignment followed the process of centroids extraction and clustering as described in Section 4.3. Regarding the affinity of tweeters to the sports betting cluster, verification that the interest per sports bettor was greater than that of the other clusters was done. This was important in the affirmation of the trained model accuracy as well as optimisation of model parameters. The sample tweets were modelled with *FastText-CBOW*, *FastText-SkipGram(SG)*, *Word2Vec-CBOW*, *Word2Vec- SkipGram(SG)* and *Glove* state-of-the-art baselines trained with $100, 200$ and 300 dimensions consistent with (Mikolov, Chen, Corrado & Dean, 2013). Models were tested with the number of clusters set to $3, 4, 5$ and 6 based on the elbow method heterogeneity measure for identification of cluster numbers befitting the dataset (Bholowalia & Kumar, 2014). The

accuracy of classifications followed a two-step process: -

- Comparative evaluation with known sports betting cluster labels in all the models, with different cluster numbers. For example, FastText's *cluster 0* represented the *Sports Betting Content*, *cluster 1* represented the *Swahili Related Chatter*, and *cluster 2* represented the *Daily News Chatter*. This was based on comparative analogy tests and manual vocabulary inspection in each of the hard clusters. The assumption was that the test set had tweets that closely inclined towards *cluster 0*, thus 0 was assumed to be the true cluster labels (ground truth). These labels were then compared with the predicted labels with different dimensions and cluster numbers. The Fowlkes-Mallows Index (FMI Score) (Rodriguez et al., 2019) was used to compute this comparison. The FMI Score is the geometric mean of the pairwise precision and recall between the true and predicted labels. The score ranges from 0 to 1. A higher value indicates better similarity between two points.

- The Silhouette Coefficient (S-Score) was computed for each model with vectors of the test tweets, assuming the ground truth unknown (Rousseeuw, 1987). The computed Silhouette Coefficient consisted of two scores: *(a) The mean distance between a sample and all other points in the same topical cluster. (b) The mean distance between a sample and all other points in the next nearest cluster.* The best value is 1 indicating the best cluster quality and the worst value is -1.

Values are reported in Table 4.2. *FastText-SkipGram* with 100 dimensions and 3 clusters, reported the highest FMI and S-Scores for the sports betting cluster. The consensus was that lower FMI and S-Scores were recorded in models with more than 3 clusters. Therefore, *FastText-SkipGram* with 100 dimensions and 3 clusters was selected to further compute *DoiSCCs*. For comparative validation against related state-of-the-art methodologies, *FastText-CBOW (100,3)*, *Word2Vec-CBOW (200,3)* and *Word2Vec-SG (100,3)* were also selected for validation based on the consistency in their **FMI** and

S-Scores.

Table 4.2: Model's classification scores with respect to (100,200,300) as model dimensions and (3,4,5 and 6) as cluster numbers.

		Fowlkes-Mallows scores (FMI), Silhouette Coefficient (S-Score)							
	#Clusters	*3*		*4*		*5*		*6*	
	Dim	**FMI Score**	**S-Score**	**FMI Score**	**S-Score**	**FMI Score**	**S-Score**	**FMI Score**	**S-Score**
FastText-Skip Gram	100	**0.65**	**0.21**	0.50	0.16	0.48	0.13	0.43	0.19
	200	0.58	0.15	0.52	0.14	0.46	0.13	0.42	0.13
	300	0.62	0.18	0.57	0.11	0.48	0.14	0.50	0.15
FastText-CBOW	100	**0.59**	**0.15**	0.52	0.16	0.53	0.15	0.43	0.18
	200	0.59	0.13	0.50	0.16	0.48	0.14	0.47	0.17
	300	0.58	0.13	0.53	0.16	0.50	0.14	0.43	0.17
Glove	100	0.57	0.12	0.53	0.12	0.45	0.11	0.45	0.12
	200	0.58	0.13	0.53	0.08	0.46	0.13	0.49	0.11
	300	0.59	0.10	0.56	0.11	0.52	0.10	0.44	0.13
Word2Vec-SkipGram	100	**0.59**	**0.15**	0.53	0.17	0.50	0.13	0.45	0.19
	200	0.58	0.13	0.50	0.14	0.45	0.11	0.43	0.12
	300	0.59	0.12	0.52	0.13	0.51	0.14	0.53	0.13
Word2Vec-CBOW	100	0.57	0.14	0.52	0.15	0.51	0.12	0.41	0.14
	200	**0.57**	**0.15**	0.51	0.14	0.47	0.12	0.42	0.13
	300	0.57	0.13	0.53	0.13	0.45	0.12	0.44	0.13
BOW (KMeans) Baseline	100 Iterations	**0.72**	**0.019**	0.59	0.015	0.48	0.016	0.53	0.016

From the values in Table 4.2, Glove models were not selected for further evaluation, because of the inconsistencies in the FMI and S-Score values. For example, the Glove model with 300 dimensions and 3 clusters had the highest FMI-Score but also the lowest S-Score for the same setting, and thus was skipped for selection. The highest and consistent values in each modelling framework are highlighted in the table. In addition, the above results meant that the dataset's best representation was three topics, especially with the consistency in the reduction of S-Score across the models and cluster numbers. For the baseline, a Bag-of-Words (BOW) K-means based clustering algorithm was used (Arthur & Vassilvitskii, 2006). The maximum iterations were 100 across the dataset. As much as the FMI score was quite high compared to the other models, the S-Scores were quite low across the clusters compared to the other models. This was an indication of a subpar performance in terms of cluster quality. Therefore, the models with the highest and consistent FMI and S-Scores across the clusters were selected for further evaluation.

4.4.3 Test Set Collection and Computation

Computation of $DoiSCCs$ for sample tweeters in Section 4.4.1 entailed the collection of tweets disseminated by the same tweeters. The intuition in this experiment was to have generic tweeters who fit this profiling. A maximum of 3200 tweets from each tweeter were collected via Twitter's search API. The numbers varied depending on how many tweets individual users had disseminated over the period of interest. Tweets from 200 tweeters were sampled and collected between 1/1/2019 to 1/04/2019. The assumption was that there was a possibility of identifying generic tweeters who could have been disseminating content related to the topics of interest in this study.

Homophily in Tweeter's Friendship Network

Analysis of a tweeter's friendship network (*mentions, retweets, replies, lists* and *hashtags*) helps identify whether the presented tweeter's profile was relevant in terms of positive correlation with the $DoiSCCs$ of his/her friends. Homophily (McPherson et al., 2001) is evident in social networks based on the fact that users tend to follow those whom they share interests with. 62275 tweets from the timelines of tweeters who bidirectionally retweeted, listed, followed, and were mentioned by the 200 tweeters as detailed in Section 4.4.1 were extracted. This was to compute their $DoiSCCs$ and correlation with the $DoiSCCs$ of their friendship network. The assumption was that a positive correlation to a large extent indicated better performance in the chosen modelling framework in terms of profiling tweeters with respect to the generated topics of interest, thus their topical affinity.

4.4.4 Parameter Settings and Experiments

The optimised *FastText-SkipGram* model had the following parameters setup: *size = 100, minimum count = 2, learning rate (lr) = 0.1* and *iteration (iter) =30*. Minimum

count of characters was two. Words with lesser character counts were excluded in the modelling as they were likely to be stop words. Descriptions of the above parameters are given in Section 4.3. Other parameters that were not explicitly specified, assumed *FastText's* and *Word2Vec's* default parameters. The model outputs were vector representations of vocabulary in the 62275 tweets. The number of clusters as well as the initialisation mechanism via k-means++ was specified in the clustering process to generate cluster centroid maps. Basically, the centroid maps are a list of terms in the dictionary embedded with their respective cluster indexes.

The training corpus in Section 4.3 was used to model tweets and tweeters. The most representative number of clusters in the chosen modelling framework, i.e., *FastText-SG* with 100 was 3 as per the results in Table 4.2. Each of the clusters had a unique user-defined identifier. *Cluster 0* represented the *Sports Betting Content*, *Cluster 1*, *Swahili Related Chatter*, and *Cluster 2*, *Daily News Chatter* (**DoiSCC**). This was after manual inspection of the vocabulary under each cluster as well as results from test terms analogy tests on the model. One assumption relating to *Cluster 0*, i.e., *Sports Betting Content* was that this cluster was expected as content relevant to this domain was appended to the generic dataset as ground truth for validation. However, *Cluster 1*, *Swahili Related Chatter*, and *Cluster 2*, *Daily News Chatter* clusters were generalised from the dataset by the modelling framework.

The resultant vectors from the modelling frameworks were then used to compute the tweet-cluster similarity. The similarity as pointed out earlier is the distance between the average tweet vector and the centroid map of interest. This process on a sample tweet is illustrated below: -

- **Original tweet** - *Away Win 3 Multibet Football Tips Odds Kenya January 11 2019 http://www.zuribet.com/away-win-3-multibet-football-tips-odds-kenya-january-11-2019/pic.twitter.com /1at2nLy8je*

- **Preprocessed Tweet** - [*away, multibet, football, tips, odds, kenya, january*]
- **Cluster Similarity values** - [**0.496**,0.196,0.434] where the value in the array index 0 represents the similarity of the tweet to the sports betting related cluster. The second value is in relation to *Swahili related chatter* and the third, *daily news chatter*. From the above example, the modelled tweet had seven terms. Kenya was a dominant term in *Cluster 1* (*Swahili related chatter*) justifying the score of 0.196. On the other hand, words like *football* and *away* were present in clusters with daily news updates and betting. This justifies the closeness in their similarities scores.

From the output, it can inferred that the above tweet was more closely related to the sports betting domain than the other two clusters.

4.5 Results

4.5.1 Group Recommendations

Social media users tend to have ties with their friends based on follower-followee relationships. Therefore, the assumption is that users with such relationships tend to share interests and by extension $DoiSCC$ values for both groups. This means that user $DoiSCCs$ for tweeters and their friendship network correlate.

In this evaluation, group topical recommendations were of choice as opposed to individual analyses. Grouping and correlating $DoiSCCs$ for both tweeters and their friendship network were the best way to represent such recommendations. Ideally, it's easier to make recommendations over a large spectrum of users as compared to individuals. For example, users with 10 to 15 percent interest in certain content, would better be grouped together as their interest similarities closely match. Each modelling

framework's vector representation differed in the computation of topical affinity for users in the three topics of interest. Therefore, $DoiSCC$ values also differed.

The overall distribution per model for the $DoiSCC$ values was computed to inform group-level recommendations. The values varied per topic simply because interest in some topical clusters, e.g., *daily news chatter* was greater than the *Swahili chatter* across the modelling frameworks. Density distributions for each of the model's $DoiSCCs$ in each topical cluster were computed to inform the grouping process as shown in Table 4.3. For example, with FastText-SkipGram (100,3), interest in *sports betting* content ranged from 0 to 0.3 while in *daily news chatter*, it ranged between 0.4 to 0.8. Inasmuch as the dataset was enriched with *sports betting* related tweets, a vast majority of tweets depicted a higher similarity to the *daily news chatter* topical cluster. This can be quantified by observing tweeting patterns where tweets mostly related to daily news were shared. One other reason is that the vocabulary in the two cluster centroids were also contextually shared. The latter values meant that the model was able to accurately identify users with interest in daily news. However, interest in *Swahili related chatter* was quite low, i.e., between 0 to 0.2 across the modelling frameworks. The dashes (-) in Table 4.3 meant that the model did not find $DoiSCC$ for that specific topical cluster. For example, *FastText-CBOW (100,3)* and *Word2Vec-SG (100,3)* could not positively identify *Swahili related* topical content resulting in null values in the interest groups.

Table 4.3: $DoiSCCs$ with respect to the modelling frameworks and topical clusters in the dataset

Modelling Framework	**FastText- CBOW (100,3)**			**FastText - SG (100,3)**			**Word2Vec - SG (100,3)**			**Word2Vec - CBOW (200,3)**			**BOW - KMeans**		
Topical Cluster	***Sports Betting (DoiSB)***	***Swahili Related Chatter (DoiSRC)***	***Daily News Chatter (DoiDNC)***	***Sports Betting (DoiSB)***	***Swahili Related Chatter (DoiSRC)***	***Daily News Chatter (DoiDNC)***	***Sports Betting (DoiSB)***	***Swahili Related Chatter (DoiSRC)***	***Daily News Chatter (DoiDNC)***	***Sports Betting (DoiSB)***	***Swahili Related Chatter (DoiSRC)***	***Daily News Chatter (DoiDNC)***	***Sports Betting (DoiSB)***	***Swahili Related Chatter (DoiSRC)***	***Daily News Chatter (DoiDNC)***
Group 1	0.0	-	0.0 - 0.1	0.0	0.0 - 0.1	0.4 - 0.5	0.0	-	0.4 - 0.5	0.0	0.0 - 0.1	0.0 - 0.1	0.0 - 0.5	0.8 - 0.9	0.0 - 0.1
Group 2	0.0 - 0.1	-	0.1 - 0.2	0.0 - 0.1	0.1 - 0.2	0.5 - 0.6	0.0 - 0.1	-	0.5 - 0.6	0.0 - 0.1	0.1 - 0.2	0.1 - 0.2	0.5 - 0.7	-	0.1 - 0.3
Group 3	0.1 - 0.2	-	0.2 - 0.3	0.1 - 0.2	-	0.6 - 0.7	0.1 - 0.2	-	-	0.1 - 0.2	-	0.2 - 0.3	0.7 - 0.9	-	0.3 - 0.5
Group 4	0.2 - 0.3	-	0.3 - 0.4	0.2 - 0.3	-	0.7 - 0.8	-	-	-	0.2 - 0.3	-	0.3 - 0.4	-	-	0.5 - 0.7
Group 5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.7 - 0.9

4.5.2 Follow-Back Recommendations for Tweeters vs Friendship Network

Following the theory of homophily, tweeters and their friendship network $DoiSCCs$ should ideally be in the same semantic space. The $DoiSCCs$ extraction process follows the same procedure as in Section 4.3.3. In the setup in Table 4.3, interest computation between tweeters and their friendship network was in relation to the three topical clusters. The three topics of interest were the most optimal in the dataset as per the FMI and S-Scores in Table 4.2 thus were not manually selected.

Therefore, follow-back recommendations were pegged on the group relations between tweeters and their friendship network in terms of $DoiSCCs$. This details the correlation of group $DoiSCCs$ in both tweeters and their network to the three clusters of interest, as in this case. The DoI generation process is synonymous to the one in Section 4.3.4. To analyse these correlations better, intra-user $DoiSCCs$ correlations in relation to the topical clusters in sports betting, Swahili related chatter and daily news are presented.

Sports Betting Topical Cluster Interest *(DoiSB)*

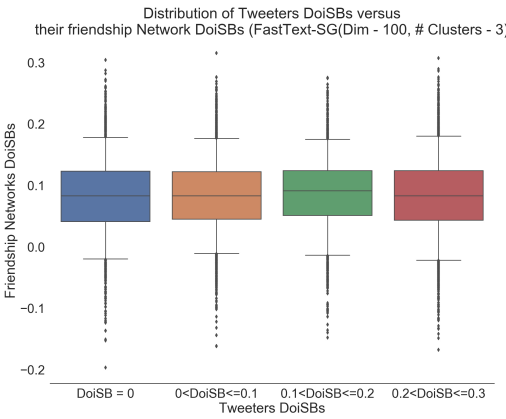
Results in Figure 4.3 show the correlation distribution between the $DoiSBs$ of tweeters and their friendship network in the models. A comparative evaluation was made with four other modelling frameworks for validation. From the box plots in Figure 4.3, correlations between the different groups' $DoiSBs$ showed marginal variance to those of their friendship network. In the case of *FastText-SkipGram (100,3)*, tweeters with $DoiSBs$ equal to 0 in Figure 4.3 (a) correlated with friends whose $DoiSBs$ median was approximately 0.08. The same pattern can be observed for the second group, i.e., with tweeters' $DoiSBs$ greater than 0 and less than or equal to 0.1.

Tweeters with $DoiSBs$ between 0.1 and 0.2 correlated better with friends whose

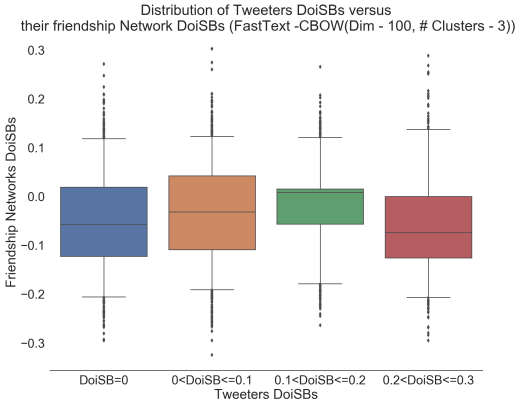
$DoiSBs$ were about 0.09. Considering the last group, i.e., tweeters with $DoiSBs$ between 0.2 and 0.3 were expected to show higher friendship correlation values. However, their best set of friendships were the same as those in the first and second group, i.e., with $DoiSB$ = 0.08. This was attributed to the number of tweeters in this specific group. Generally, they were fewer than Group 3's in the dataset. This further influenced their friendship connections, which explains Group 4 results in Figure 4.3 (a). Results in Figure 4.3(e) were a bit inconsistent with vector representations, as the algorithm is purely based on the bag-of-words approach. This explains why the $DoiSBs$ are quite high across the groups. These results may look great but the fact that no semantics are factored in the model, overfitting in the word re-usage across the classifications becomes a problem. This points to the inconsistencies in Table 4.3 when humans evaluate the same model and, in the FMI-Scores for BOW in Table 4.2. Therefore, inasmuch as the model exhibited positivity in correlations, the same could not be justified when humans validated the same model. In addition, the rest of the other models except *FastText-SkipGram (100,3)* depicted negative correlations between tweeters' and their friendship networks' $DoiSBs$ as much as they were contextually similar. This shows that although some tweeters had interest in online sports betting, their friends did not, which contradicted the *homophily theory*. This can also be partly attributed to the model quality which is evident in Table 4.2 for the models in Figures 4.3 (b), (c) and (d). Figure 4.3(a) shows otherwise.

Swahili Related Chatter Topical Cluster Interest (*(DoiSRCs)*

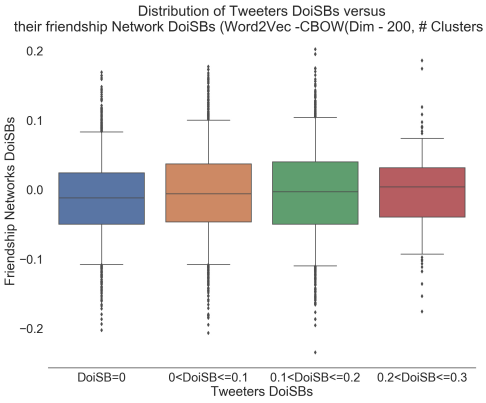
Results in Figure 4.4 depict the same intra-user correlations with regard to the Swahili Related Chatter ($DoiSRC$) topical cluster. From the results, only modelling frameworks based on *FastText-SkipGram (100,3)* and *Word2Vec-CBOW (200,3)* depicted positive correlation in their $DoiSRCs$. *Word2Vec-SkipGram (100,3)*, *FastText-CBOW (200,3)* and the Bag-of-Words (BOW) baseline frameworks could not entirely model



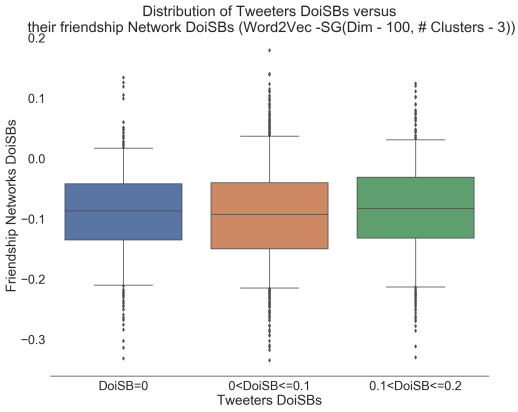
(a) FastText-SkipGram (100,3)



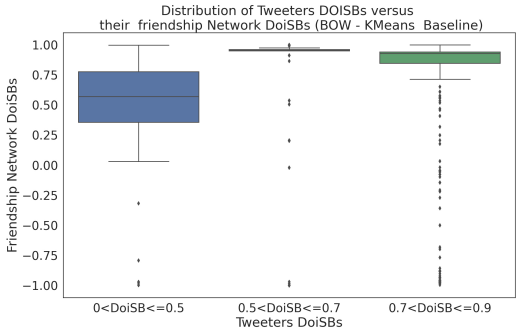
(b) FastText-CBOW (100,3)



(c) Word2Vec-CBOW (200,3)



(d) Word2Vec-SkipGram (100,3)

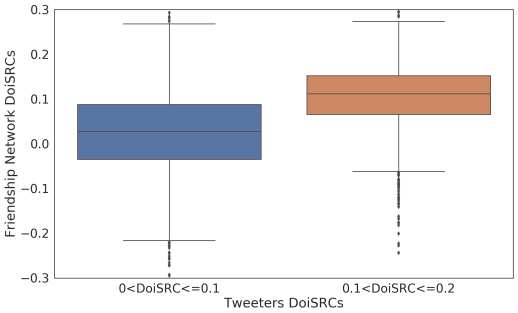


(e) BOW-KMeans Baseline

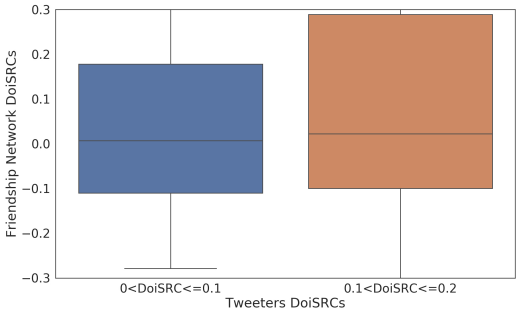
Figure 4.3: Correlation of tweeters and their friendship network Degree of Interest in Sports Betting (DoiSB) in four models

any positive correlations in relation to the *Swahili Related Chatter* topical cluster as evidenced in Table 4.3. In Figure 4.3, *FastText-SkipGram (100,3)* depicted the most

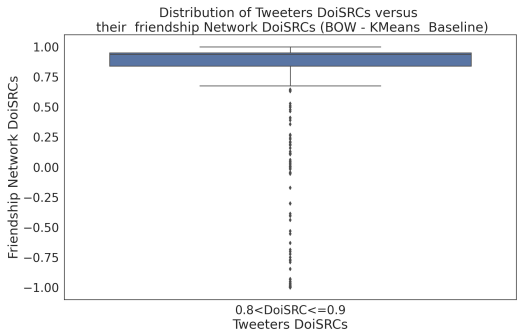
positive correlation in the values thus was the best performing framework in the identification of the tweeters and their friendship network propagating Swahili related content.



(a) FastText-SkipGram (100,3)



(b) Word2Vec-CBOW(200,3)



(c) BOW-KMeans Baseline

Figure 4.4: Correlation of tweeters and their friendship network Degree of Interest in Swahili Related Chatter (DoiSRC)

Daily News Chatter Topical Cluster Interest (*DoiDNC*)

The Degree of Interest in *Daily News Chatter* ($DoiDNC$) in relation to the third and final topical cluster was computed. This cluster entailed users who disseminated content related to daily news over the collection period. Due to the volatile nature of content dissemination patterns on short-text microblogs, citizen journalism has been on the rise. Tweeters share daily news items with their networks, forming a large chunk of these news related topical content clusters. Since the *Daily News Chatter* was one of the extracted generic clusters from the vector representation of the dataset, there was a need

to compute the interest in the same. The aim was to correlate the interest of users and their friendship network in news related content.

From the results in Figure 4.5, *FastText-SkipGram (100,3) (4a)* modelling framework extracted and correlated the interests better than the other three models shown in the figure. $DoiDNC$ values were the highest of all the modelling frameworks as well as the depiction of better correlations in both tweeters and friendship network. *Word2Vec-SkipGram (100,3)* only identified $DoiDNCs$ falling in two categories as evidenced in Table 4.3. On the other hand, *FastText-CBOW (100,3) (4b)* and *Word2Vec-CBOW (200,3) (4c)* had the lowest $DoiDNCs$ for both tweeters and their friendship network.

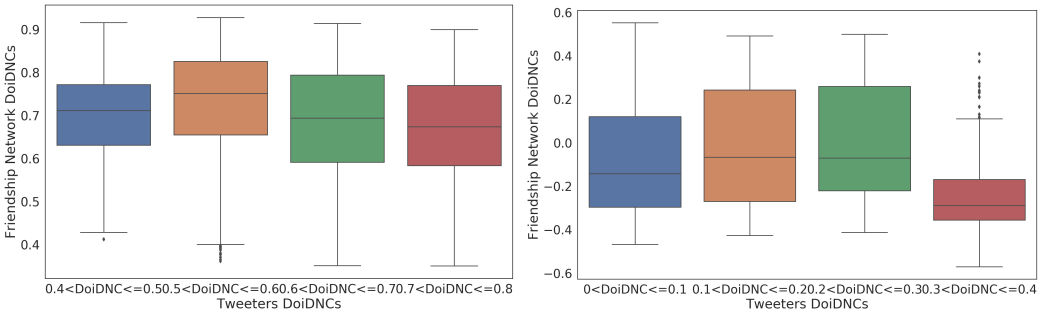
Evaluations in sports betting, Swahili chatter and daily news chatter sections, justify the choice of the modelling framework. *FastText-SkipGram (100,3)* framework depicted the best performance among vector representation techniques as well as when compared to the *Bag-of-Words* baseline.

4.5.3 Qualitative Evaluation in Follow-Back Recommendations

Homophily is the tendency of users to have positive ties with other similar users in socially significant ways. Currently, the term is used to refer to an observable behavioural tendency whose causes can include preference or opportunity (van den Beukel, Goos & Treur, 2019).

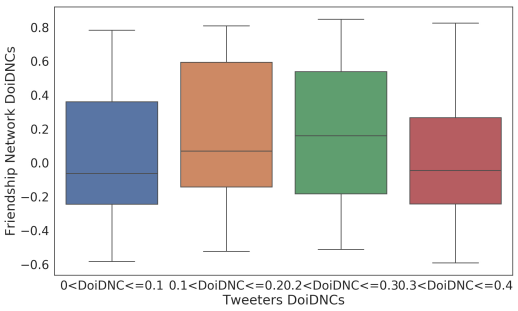
Homophily in Short-Text Microblog Users

To evaluate the results qualitatively, a *DoiSCC* topical classification group that was most consistent across the modelling frameworks was selected. Group 3 ($0.1 < DoiSBs <= 0.2$) in Table 4.3 under the *sports betting* topical cluster fit this criterion. Ground truth tweets in the sports betting domain were also added earlier to the dataset as in Section 4.4.1. This meant that users with 0.1 to 0.2 interest in the sports betting domain could be

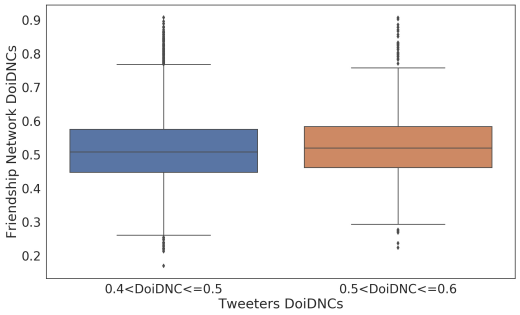


(a) FastText-SkipGram (100,3)

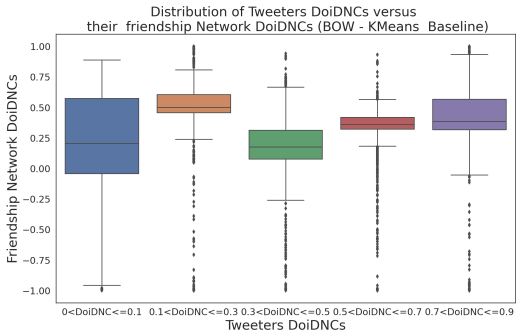
(b) FastText-CBOW (100,3)



(c) Word2Vec-CBOW (200,3)



(d) Word2Vec-SkipGram (100,3)



(e) BOW - KMeans Baseline

Figure 4.5: Correlation of tweeters and their friendship network Degree of Interest in Daily News Chatter (DoiDNC)

identified in the dataset despite the modelling algorithm. This is partly attributed to the introduction of the ground truth sample data as described in Section 4.4.2. The process of selecting tweeter's and their friendship network tweets for *DoiSBs* computation and correlation followed the process outlined below: -

1. Three random users with $0.1 < DoiSBs <= 0.2$ and who disseminated at least

30 tweets in the initial collection were selected for further evaluation. To the best of our knowledge, 30 tweets could somewhat be modelled thus the user's interest level in a certain topical cluster could be computed. Tweets collected were recorded under each of their usernames and had to be in English. Non-English ones were removed from the set.

2. A collection of English tweets from the user's friendship network's list, i.e., those who *"mentioned","retweeted", or "replied"* to tweets disseminated by the tweeters as in point 1. One other condition was that the users in the friendship network should have disseminated at least 30 English tweets too. A minimum of 30 tweets in friends and their network provided enough data for modelling, as it depicted better online user activity.

3. The qualitative evaluation process centred around one topical cluster, i.e., the *sports betting* one as mentioned earlier. Therefore, tweets were classified in either the *betting* or *others* classes for consistency. This meant that any tweet that did not depict any betting related content as per human understanding was placed in the *others* class.

On the other hand, evaluators were selected to specifically look at the semantic correlation between tweeters and their friendship network's tweets in relation to the *sports betting* topical cluster. The *sports betting* cluster was chosen as it represented the ground truth as detailed in Section 4.4.2 thus labelling the tweets was faster and more accurate. The semantic correlation in the friendship network was averaged across the evaluators. To ascertain this, the level of agreement in terms of the semantic relation between user's tweets and their friendship network's was computed. The three evaluators were selected based on their knowledge of the English language and familiarity with betting related terms. They followed a three-step process in assigning tweets to respective topical clusters: -

1. Firstly, the evaluators were presented with a list of 100 tweets from the *sports betting* set as described in Section 4.4.1. They were required to go through the list at least three times for a deeper contextual interpretation of tweet content in this domain.

2. Secondly, the evaluators were required to label 30 random tweets assigned to them from the tweeters and 30 from respective friendship network. The assignment was in either the *sports betting* or *others* topical clusters. The assumption was that, once an evaluator picked a tweeter and his/her tweets, he/she was also obliged to classify tweets of that specific tweeter's friendship network in the presented list. This further provided a deeper contextual understanding of the disseminated content between tweeter's and their friendship network.

3. Finally, the inter-evaluator agreement across the *sports betting* topical cluster in form of Cohen Kappa scores (McHugh, 2012) was computed. A higher inter-evaluator agreement value meant more relevant friendship networks. Primarily, this process was the final affirmation of *homophily* among users on short-text microblogs validating the follower-followee relationships. It proved that topical affinity in users could not only be measured by the disseminated content, but also by analysis of the friendship network content.

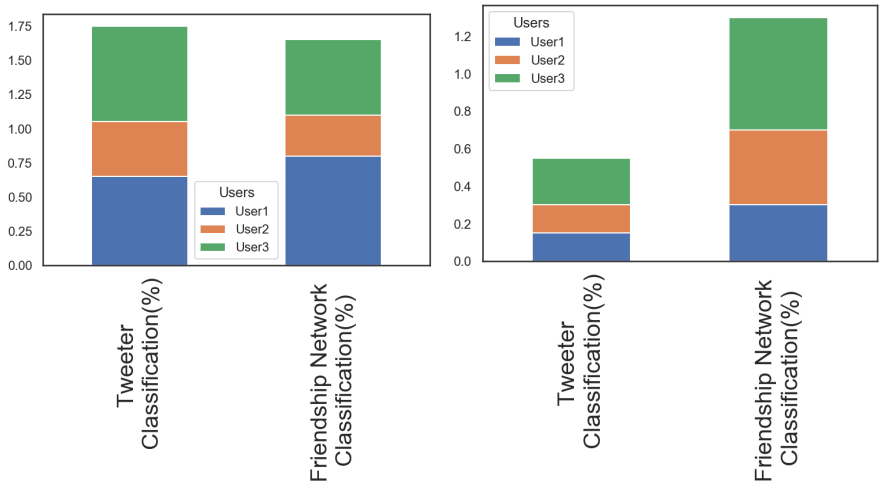
Figure 4.6 shows the classification accuracies with respect to *online sports betting* for three randomised users. The values on the y-axis are the percentages on a scale of 0 to 1. Consistency and percentage of accuracies was the measure of qualitative performance for tweeters as well as their friendship network. Across the selected tweeters and their friendship network, ***FastText-SkipGram (100,3)*** consistently performed well in identification of sports betting related content as depicted in Figure 4.6 (a). For example, for one random user (labelled User1 in the graph), his/her classification results for tweeters were 65% while those of their friendship network were approximately 75%, a

strong indicator of users with shared interests. The same was replicated for another user (labelled User3 in the graph) disseminating approximately 70% of sports betting related content with around 60% of his/her friendship network disseminating related content.

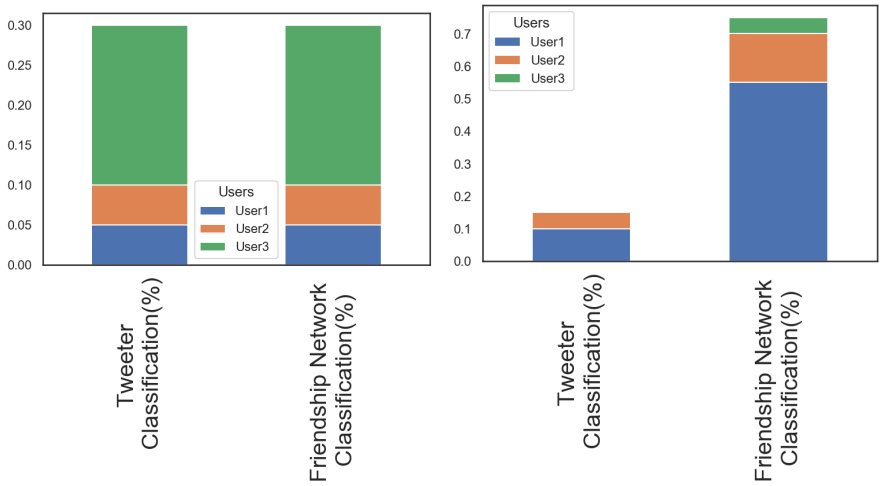
Word2Vec-CBOW (200,3) had the lowest accuracies in identification of sports betting related content in tweeters and their friendship network as shown in Figure 4.6 (c). *User 1* in this framework recorded very minimal interest in sports betting, i.e., about 4% while his/her friendship network recorded a 5% interest in such content. On the other hand, ***FastText-CBOW (100,3)*** in Figure 4.6 (b) consistently depicted better classification accuracies compared to ***Word2Vec-SkipGram (100,3)*** (Figure 4.6 (d)). The *Bag-of-Words* baseline results were also inconsistent with the *homophily theory*. For example, evaluation by *User3* placed sports betting related tweets at 50% sports betting-related content. The comparative friendship network disseminated just 21% sports betting related content. These results corroborate quantitative findings in Section 4.3.4 where ***FastText-SkipGram (100,3)*** depicted the best **FMI** and **S-Scores**. Generally, *FastText* makes use of a sliding window of characters making it relevant in datasets with shortened or misspelled words such as in tweets. This is a plausible conclusion as to why *FastText* based models performed better.

Semantic Correlations

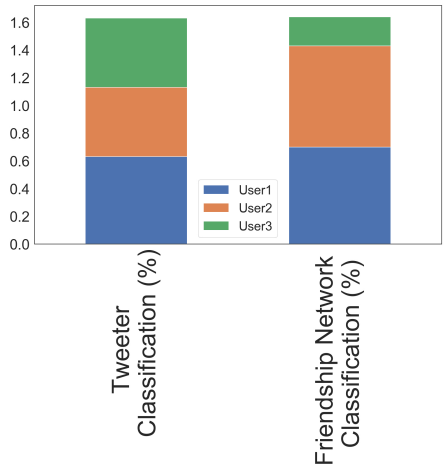
In addition to the inter-rater agreements in evaluation of homophily in Section 4.5.3, an evaluation of the semantic correlation between the selected users and their friendship network tweets was performed. The assumption was that for the two sets to be semantically similar, their representations should ideally be in the same semantic space. Therefore, the Pearson Correlation Coefficient (PCC) between the sets of user tweets and their friendship network was computed (Benesty, Chen, Huang & Cohen, 2009). There is an assumption of linearity of users and their friendship network in the same semantic space. For example, users with $DoiSB$ or $DoiDNC$ values between 0 and



(a) FastText-SkipGram (100,3) (b) FastText-CBOW (100,3)



(c) Word2Vec-CBOW (200,3) (d) Word2Vec-SkipGram (100,3)



(e) BoW (K-Means) Baseline

Figure 4.6: Qualitative Evaluation of Sports Betting Content Classification Accuracy for Tweeters and their Friendship Network

0.3 should ideally have values in the same range in their friendship networks. Therefore, a correlation between $DoiDNCs$ of the two groups assumes linearity at least to a certain level. The PCC is then computed as the covariance of the two variables per classification groups of $DOIs$ as indicated in Section 4.5.1 divided by the product of their standard deviations. This is represented as below: -

$$PCC = \mathrm{Cov}(a, b)/\sigma(a) * \sigma(b) \tag{4.5}$$

σ is the standard deviation that is applied to variables a and b. The correlation is a value between -1 and 1, where 1 is an indicator of positive correlation, -1, negative correlation and 0 indicating no correlation. In the computation of $PCCs$ to validate follow-back recommendations, specific groups were selected from both sets of classifications. For example, if users in classification $0.7 < DoiDNC <= 0.8$ are selected, then the comparison is only made with their friendship network $DoiDNCs$ in the same classification group. A perfect scenario would be almost equal values in the comparative groups.

In Table 4.4, it is evident that the *Bag-of-Words* baseline results are dismal in correlating what users and their friendship networks disseminated. This output corroborates the results in the classification scores in Table 4.2. *FastText-SkipGram (100,3)* performed the best in identification of semantic correlations between users and their friendship networks with a PCC of 0.7 in the classification group. The choices of the classification group of interest and related interest range was based on the results in Table 4.3. $DoiDNCs$ were uniformly identified across the modelling frameworks as this was the most optimal for evaluation. The interest range was one with the highest values in the group, i.e., the most representative of the interest and had at least 30 users and related friendship network. The only exception was with *FastText-CBOW (100,3)* where the lowest interest value was selected as no other interest range fulfilled the

Table 4.4: Follow-back correlations as Pearson Correlation Coefficients (PCCs) between users and their friendship networks in with interest in Daily News Chatter classification group

Modelling Framework	**Degree of Interest in Daily News Chatter (DoiDNC) Classification Group**	**Pearson Correlation Coefficient**
FastText CBOW (100,3)	0.0 - 0.1	0.657
FastText SkipGram (100,3)	0.7 - 0.8	**0.700**
Word2Vec SkipGram (100,3)	0.5 - 0.6	0.106
Word2Vec CBOW (200,3)	0.3 - 0.4	0.500
Bag of Words - KMeans Baseline	0.7 - 0.9	-0.110

30-user count in both users and friendship network except this range. The comparisons were based on the most representative users in the lowest range and in the same interest group. Such a measure is statistically significant in this evaluation.

4.5.4 Application Areas

Quantitative and qualitative results obtained in Sections 4.5 affirm the possibility of extracting relevant user-representative interests in modelling topical clusters in short-text microblog content. Experimental processes in this setup affirmed the choice for a FastText-based modelling framework as the ideal modelling framework in such scenarios. However, the same modelling process is applicable in several other ways on short-text microblogs as below: -

- **Follower - Followee recommendations** - User interactions on short-text microblogs help in content propagation on such platforms. Content in terms of *hashtags*, *"trend"* based on the rate at which users on the platforms re-share the content. Users with interest in certain content should therefore be in the same semantic space. Basing the argument on the ***FastText-SkipGram (100,3)*** (Figure 4.3 (a)), tweeters with $0.1 < DoiSBs <= 0.2$ can be recommended for follow-back to users

whose $DoiSBs >= 0.09$. This helps in building more solid semantic connections which ultimately propagates content to a wider and relevant audience.

- **Cold-start recommendations** - New users on microblog platforms struggle to find users with aligned interests due to lack of content and sometimes irrelevant connection suggestions. Based on the same FastText model, tweeters neutral to sports betting, i.e, $DoiSBs = 0$ tend to correlate well with friendship networks with $DoiSBs >= 0.0.08$. Factoring in other attributes such as location metadata, such tweeters are likely to be recommended to the cold-start users for follow-back.

- **Third-party content propagation** - Content-based recommender systems are reinforced by relevant shared content over time. Therefore, with better computation of user interest levels in certain topics, more semantically relevant third-party are bound to be recommended to users with varied $DoiSCCs$.

4.6 Chapter Summary

Users on the short-text microblog platforms depict preference towards several topics to a lesser or greater extent. This is influenced by their online interactions, i.e., based on the interests of other tweeters or largely by personal preferences. The interactions form relationships among tweeters. The relationships may either be unidirectional or bi-directional. Therefore, such tweeters need support of recommender systems to identify the right tweeter to follow back or content to propagate. In this chapter, an approach that can be used to learn the semantic relevance of tweets was introduced. The new approach identifies the degree of user representative interests a tweeter has in a topic considering the semantic relevance of the user's tweets. A set of neural network-based models was considered in accomplishing this task. In the experiments, a *FastText* model with 100 dimensions and a little bit of hyper-parameter tuning worked

best in the extraction of words with high semantic relevance to the generated topical clusters. To do this, input corpus for model training and vector representation were modelled. Clustering on the corpus to extract the centroids maps representative of topical clusters was also computed. These centroid maps were then used to determine the overall closeness of tweets to the generated clusters.

Results in the chosen model (*FastText*) corroborated the *homophily theory* whereby users who were highly involved in disseminating, for example, sports betting related tweets, had friends who disseminated related content. This supports the argument stated in this chapter which also aligns with the fundamental principal in follower-followee based social networks. In the experiments, thresholds denoted by the interest groups were determined to define the DoI. These DoI values can be used for the design of recommender systems in this domain.

User-representative interests extraction methodology was developed in a way that the semantic distance to a certain interest cluster is computed. A follow-back mechanism was also developed to complement the interest identification process. To further advance this approach, a multi-interest approach in generating more user-representative interests was developed in Chapter 5.

Chapter 5

Multi-Interest Modelling

5.1 Introduction

This chapter extends the work presented in Chapter 4, where user-representative profiles were generated by computing the distance between user generated content and dataset representative interest clusters in short-text microblogs.

Tweeters are defined as users who disseminate content on Twitter, for example, extrinsically declaring their interests at sign up. Changes to these original interests are still user driven which is not always the case for many short-text microblog users. Such changes are in the form of new friendships or *"hashtag"* suggestions, which do not overly represent the true identities of such online users. Diversity in the disseminated content makes it difficult to align interests of one user to those they declared at sign up. For instance, a tweeter may continuously disseminate political related content during electioneering times, but also occasionally tweet about his favourite football team yet his extrinsic profile leans towards car racing.

5.1.1 Notations and Symbols

The below notations and respective descriptions are used in this chapter: -

x - User, subject to extraction of their profile.

Y_x - Rows of user x tweets.

Z_w - Summation of user interest vectors per topic.

E - User representative model.

$MiUP$ - Multi-interest User Profile (Average interest values per user, across the dataset).

$x_{1u}..x_{3u}$ - Evaluator soft topic classifications.

$M_{1u}..M_{3u}$ - Model classification of a tweet.

$k_{1u}..k_{3u}$ - Kappa scores across topics.

p_e - Hypothetical probability of chance agreement between the model's and evaluators' observations.

p_o - Relative observed agreement between the model's and evaluators' observations.

5.2 Problem Statement

Tweeters for example extrinsically declare their interests at sign up. Changes to these original interests are still user driven which is not always the case for many short-text microblog users. Such changes are in the form of new friendships or *"hashtag"* suggestions which do not overly represent the true identities of such online users. Diversity in the disseminated content makes it difficult to align interests of one user to the declared ones at sign up. For instance, a tweeter may continuously disseminate politically related content during electioneering times, but also occasionally tweet about his favourite football team, yet his extrinsic profile leans towards car racing. Capturing

the diverse nature of interests in this respect for better presentation of third-party recommendations, e.g., for personalised marketing, is important. Without capturing the context, keywords are insufficient in most cases as they may be used in negation. For example, a tweet such as *"I don't like KFC"* depicts a user with extrinsic lack of interest in KFC. However, keyword search by the word "KFC" might just present this as a potential entity of interest yet has been declared not to be. The same argument in political conversations, though negative, may still be of interest as a change of opinion is possible. Therefore, contextual understanding of short texts in the extraction of varied user interests for diversification in user profiling is paramount. In formulating this modelling process, the following research questions arise : -

- Are user representative interests deducible in short-text microblogs, based solely on disseminated content?
- Is it possible to extract diverse but user-representative topics of interests in streaming short-text microblogs?
- Is it possible to quantify a user's level of interest in various topics based on the disseminated content?

In this chapter, a framework that extracts diverse topics of interest and computes user's interests based on the disseminated content is presented. This is an augmentation to the approach in Chapter 4. Users on short-text microblogging platforms present divergent spaces in terms of topical interests. A responsibility matrix depicting users and their levels of interest in the varied topics which ultimately defines their profiles is computed. To validate the framework performance, a generic Twitter dataset geolocated to Kenya over a period of one year is considered. Kenyan Twitterspace is considered based on the author's knowledge of tweeting patterns and language in the country as well as the diversity in topics over the time.

As mentioned above, this work is a build up to the initial work in topical affinity in short texts in Chapter 4. The difference with the earlier work is in terms of responsibility matrix computational aspect as well as the approach used in the deduction of diverse topics of interest in defining user representative profiles. Contributions focused in this chapter are as below: -

1. Design of a user profiling framework that considers the most definitive aspect of online users in true extrinsic identity deduction, i.e., the disseminated content.

2. Development of a soft computation approach that assigns an interest responsibility level per topic to each user. This ultimately generates more succinct profiles.

3. Testing of the framework using known Twitter users in the Kenyan Twitterspace. Validation of the output in the formulation of true user identities of such short-text microblog users is also computed.

4. In the design science aspect, the work encompasses the definition of the research problem and development of artefacts (Grenha Teixeira et al., 2017), (Lapão, Da Silva & Gregório, 2017). This is done through modelling via neural networks approaches. The output in the form of word embeddings is the input to the Gaussian Mixture Model (GMM) with Expectation Maximisation (EM). EM builds soft clusters to compute the degree of interest that users have in certain topics in the dataset. The processes are detailed to the quantifiable outputs and validation by human evaluators to make sure that the model works, and objectives are met.

5.3 Multi-Interest Modelling Approach and Summary of Literature

The core profiling processes in the proposed framework are detailed. The framework encompasses processes related to *modelling and representation of short texts*, *clustering,* and *user interest-based responsibility matrix computation* [1]

Therefore, several neural network approaches were adopted in the representation of tweets. This assumption is based on the success of other neural network approaches in short texts as demonstrated in (J. Li et al., 2016), (Mishra et al., 2018), (Zhang et al., 2018). Approaches related to Expectation Maximization(EM) to compute topical responsibilities as well as the choice of cluster centres as interests, were made use of (Dempster, Laird & Rubin, 1977; Arthur & Vassilvitskii, 2007). A more detailed review of literature is found in Section 2.5. As mentioned earlier in the chapter, this work precedes the research on the determination of the degree of interest a user is likely to have in a certain topic as explained in Section 4. Figure 5.1 summarises the steps in the multi-interest profiling process as listed above. Outputs of each step are embedded.

The following steps were followed in computing user-to-topic responsibilities in the dataset. The topics were represented as interest clusters.

5.3.1 Modelling Short Texts

The modelling process was based on *Word2Vec, FastText* and *Glove* neural network algorithms. The optimised performance among them was quite close, based on the classification performance over other neural network approaches as depicted in Figure

[1]This work (Wandabwa, Naeem, Mirza, Pears & Nguyen, 2020) was submitted, and accepted to the 15th International Conference on Design Science Research in Information Systems and Technology (DESRIST). Published in Designing for Digital Transformation. Co-Creating Services with Citizens and Industry pp 154-168 as part of the Lecture Notes in Computer Science book series (LNCS, volume 12388).

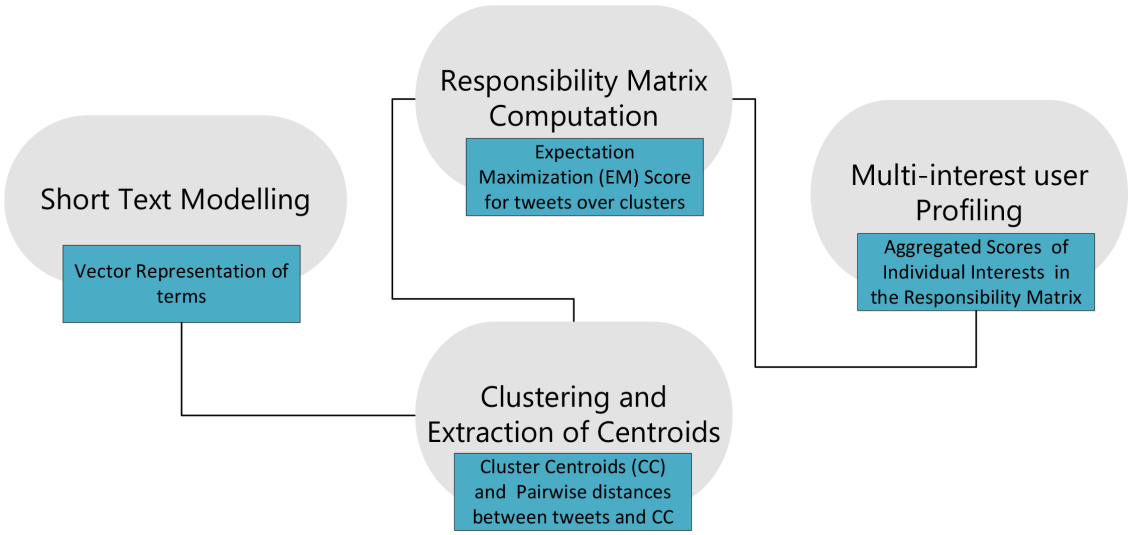


Figure 5.1: Multi- Interest Modelling Framework

5.2, presented later in this chapter. From the classification results, *Word2Vec* and *FastText* models performed well across the tests. The model outputs word embeddings that typically reconstruct linguistic contexts of words. Each word is assigned a vector space in such a way that contextually close words are located close to one another in the dimensional space (Mikolov, Sutskever et al., 2013) to factor in the co-occurrence patterns. As much as the *Word2Vec* and *FastText* based classification results were close, Word2Vec-based approaches were marginally better than FastText ones. Therefore, the modelling was done via *Word2Vec*, even with the assumption that *FastText* approaches results are still likely to be close. Cluster centroids were further extracted from the Word2Vec-based model for further experimentation as proof of concept.

Following parameters were provided in the generation of vector representations: -

- *size* or the dimensions in the vector space.
- *min_count* or *minimum count* of a word in the corpus for it to be added to the training set.
- *sg=1* for training a SkipGram model and *sg=0* for Continuous Bag of Words (CBOW). In SkipGram modelling, the algorithm uses current word in the list to

predict its neighbours (its context). However, the in CBOW, the context is used to predict the current word.

- *window* is the maximum distance between the current and predicted word in the list of word tokens.

- *iter* or iterations is the number of iterations (epochs) over the corpus throughout training.

The rest of the modelling approaches were attributed the same way as in Chapter 4. Unspecified parameters assumed the default values specific to each modelling technique. As described in Chapter 4, *FastText* vector representation ignores word structures as each word f is represented as a bag-of-character *n-grams* x. In this setup, $3 \leq n \leq 6$ was specified as in the implementation here (Bojanowski et al., 2017).

5.3.2 Clustering and Initialisation of Centroids

The intuition behind the clustering process in this case is that semantically close tweets are usually grouped together. To define clusters, manual inspection of underlying keywords in each cluster are observed. These keywords ideally are representative topics for each of the clusters. *K-Means++* was the algorithm of choice in clustering and extraction of initial centroids (Arthur & Vassilvitskii, 2007). As mentioned in Chapter 4, the algorithm initially spreads out the set of cluster centroids, thus optimising the choice of initial cluster centres. This guarantees an $O(\log k)$ solution for convergence.

The choice of the number of clusters was based on two factors. Firstly, the optimal set of topics in the dataset was determined using the conventional Latent Dirichlet Allocation (LDA) (Blei et al., 2003a). Owing to the weaknesses of modelling topics with LDA on short and sparse texts, manual seeding of the extracted topics using MELDA as described in Chapter 3 was opted. Therefore, topics that were semantically

close were merged. The final topics provided a better indication of the expected cluster numbers as ideally semantic closeness is key in both topic modelling and clustering. Secondly, a heterogeneity measure (Bholowalia & Kumar, 2014) was used to determine the optimal cluster numbers in the dataset which ideally should be as close as possible to the optimal number of topics. The well-known Elbow method (Bholowalia & Kumar, 2014) was applied in determining this. With this method, several tests were run considering different k values representing the number of clusters. To compute the measure, tests were run considering different k values with a known test set. The number of clusters that best identified the sample dataset were chosen as optimal. The cosine distance measure was then used to compute the intra-cluster distance between y points in each cluster Z_k and centroid Z_z in that cluster.

The main aim of this phase was to identify words commonly used in the cluster of interest and then compute its centroid map. Computation of centroids, albeit by word as in (Recalde & Kaskina, 2017), provided a reference point for computation of the distance between clusters (interest topics) and any other tweet.

5.3.3 Responsibility Matrix Computation

Expectation-Maximisation algorithm (Dempster et al., 1977) was applied in identification of the responsibility levels in terms of topical cluster alienation as described in Section 5.3.2. Following steps were implemented in the Expectation Maximisation (EM) computation: -

1. **Selection of initial values for estimations** - Instead of randomising the starting values, the initial set of cluster centroids were extracted based on the chosen *K*, number of topics. The cluster centroids were in vector format and extracted from the Word2Vec model. This is because Expectation Maximisation is sensitive to the initial means, and therefore a bad choice of means can lead to overlapping

points (Recalde & Baeza-Yates, 2018).

2. **Initialising of each cluster's weight based on the number of tweets in individual clusters** - The assignment of tweets to clusters was done via K-means++ to fulfil conditional expectations. The process is then followed by maximisation of the log-likelihood with respect to the cluster weights assignment where missing data points are replaced by the conditional expectation.

3. Lastly, the initial co-variance matrix is computed as convergence is assessed.

Expectation Maximisation in this case defines the extent to which topics (clusters) have influence over tweets. The resultant output after convergence is a matrix of topics and their respective responsibilities over the individual tweets. The summation of shared topical responsibilities over each tweet should add up to 1. The formulation of the above steps is in Section 5.3.4.

5.3.4 Multi-interest User Profiles

Input in the computation of multi-interest user profiles was the responsibility matrix computed in Section 5.3.3. To define a *user* x profile, *rows of individual modeled tweets* v corresponding to the user Y_x, are selected from *test dataset matrix* Y. Therefore, $Y_x \in Y$. The *user-interest representative model* is derived by averaging user interest vectors per topic. The output is representative of individual topical interests

$$Z_w = \sum_{i=0}^{|Y_x|-1} w_{xb} \tag{5.1}$$

To comprehend Equation 5.1, let w be the *count of modeled tweets per user* x. For $b \in [0, K-1]$, the user-representative interest model is computed as the average of Z_w. This is formulated as below: -

$$E = \sum_{b=0}^{K-1} Z_w \tag{5.2}$$

This is representative of the *sum of vector values* b. The Multi-interest User Profile ($MiUP$) computation is represented as below:-

$$MiUP = (Z_w/E) * 100 \tag{5.3}$$

In the formulation of Equation 5.2, b values are simply the *interest values* in each of the 10 topics averaged per user. It is essential to define users in homogeneous groups in profiling. Therefore, a threshold value n is introduced as definitive least interest value to be included in grouping users surpassing a certain interest level. The value is determined from the inter-topic interest median. For example, if the test user's median value is 0.4, then 0.4 will be set as the threshold for that topic across all users.

5.4 Experimental Framework and Setup

Processes to validate this approach are outlined in Section 5.3. For validation of the results, *Glove*, *Word2Vec* and *FastText* models are trained on the same dataset with different dimensions. A description of the source and nature of the dataset is in Section 5.4.1.

5.4.1 Datasets and Settings

The corpus to be trained comprised of $650,055$ unique tweets geo-localised to Kenya. They were collected in JSON format for a period of a year starting 17/10/2018. Each tweet entry contained associated metadata such as hashtags, mentions, links and geo-location data. All retweets were filtered from the collection. The tweeting languages were primarily English and Swahili, Kenya's national languages. Just as in Chapter 4,

the choice of Kenya's Twitterspace was largely influenced by the author's familiarity with Kenyan's tweeting patterns, diversity in topics as well as the availability of some domain specific data that augmented the dataset.

Dataset Augmentation

Regarding domain specificity and validation of this approach, the dataset was augmented with 50639 sports betting related tweets geolocalised to Kenya. The known set was important for validating the methodology. To be specific, the betting related tweets were queried and collected from Twitter handles of *sportpesa*[2], *betin*[3],*eazibet*[4], *betika*[5] and *betwayke*[6]. They are all sports betting companies with presence in Kenya. Therefore, the entire training set comprised of $700,694$ unique tweets all geolocalised to Kenya.

Sample of Users for Evaluation

To evaluate the methodology, two known but diverse datasets were made use of. 1000 tweets associated with sports betting Twitter handles labelled *Sports Betting Related* were considered. In addition, 1000 politics-related tweets generated on 8th August 2017 during the last general election in Kenya were also considered. These two divergent but geographically relevant sets of data provided a mechanism for testing the modelling approach before the profiling process. The best modelling approach should be able to separate the two sets in almost equal segments.

The 2000 tweets were then pre-processed, and duplicates filtered out. All tweets with less than 25 characters were also removed from the set. The pre-processing steps are described in detail in Chapter 4. Finally, 784 sports betting related and 769 politically-related tweets were validated.

[2]https://www.sportpesa.org/
[3]https://www.betin.co.ke/
[4]https://www.eazibet.co.ke
[5]https://www.betika.com/
[6]https://www.betway.co.ke

5.4.2 Model Training and Evaluation

The short-text modelling process described in Section 5.3.1 to train the models was followed. The trained models were based on *FastText* (Bojanowski et al., 2017), *Word2Vec* (Mikolov, Sutskever et al., 2013) and *Glove* (Pennington et al., 2014) neural network modelling techniques. The output of the modelling process is a vector representation of a word(s) based on the context within which the word is commonly used. However, for *FastText*, word modelling is independent of the language of expression and vocabulary size as it is based on a sliding window of characters instead of comparing whole words.

Model Training

Five neural network based models *Word2vec-SkipGram (100 dimensions), Word2vec-CBOW (200 dimensions)*, *FastText-SkipGram (100 dimensions), Glove (300 dimensions) and FastText-CBOW (100 dimensions)* were trained for evaluation and consistency purposes. The choice of the above models with their respective parameters was based on their success in the earlier experiments in Chapter 4. A pre-processed and tokenised corpus of tweets was the input in the model's training framework. Ideally, the models learn character or word patterns by mapping words in the corpus to vectors. Vectors that are tuples of data values representing the words are then used in the computation of word similarities. In general, the models predicted the next word based on the surrounding words, contextually. FastText models further learnt word contexts by averaging windows of character representations. These models therefore dealt better with out of vocabulary or misspelled words compared to Word2Vec or Glove based models. The models to a large extent contextually learnt Swahili and English terms well. The following parameters were specified before training *Word2Vec* and *FastText* models:

- *size* or the number of dimensions, *min_count* or least count of a word in the

corpus for it to be factored in the training set.

- The *window* parameter. This is the maximum distance between the current and predicted word in a tweet.
- *sg* for training a Continuous Bag of Words (CBOW) if $sg = 0$ or undefined and skip-gram model if $sg = 1$.
- *word n_grams* to enrich word vectors with subword (n_grams) information if specified as 1 ; and *iter* or iterations which was the number of iterations (epochs) over the corpus.
- Number of *epochs*, i.e., one full cycle in training was defined for the *Glove* model as well as the *learning rate (lr)*.

The model outputs were vectors of each word in the cleaned corpus. An in-depth view of the word representation process in the three modelling techniques is explained in Section 4.3.1 under model training.

Model Evaluation

There was need for evaluating each of the state-of-the-art modelling techniques to select the most relevant one for multi-interest user profiling approach. In this instance, each of the models is subjected to a labelled set of test data.

Based on the heterogeneity results in Figure 5.3, the dataset had approximately 22 topics. Out of the 22, there was ground truth tweets for two topics, i.e., ***Sports Betting*** and ***Politics***. The tweets were extracted from timelines of betting companies and politicians in Kenya respectively which were then aggregated with the training data. Each tweet was annotated as *political* or *sports betting* related based on the disseminating Twitter handle. The heterogeneity measurement is computed via *K-means++* as illustrated in Section 5.5.

5.5 Results

5.5.1 Topical Classifications

A scrutiny of tweets especially in the list of hashtagged ones, gave an approximation on the number of topics. In addition, they also provided better generalisation, especially with the smaller topics. Hashtags such as *#uhurumustgonow*, *#IstandwithNdindiNyoro*, *#PunguzaMizigoBill2019* were labelled as ***Political Discourse***. On the other hand, hashtags such as *#mensfinal, #USOpen, #Nadal* define the ***Sports*** topic. Tweets classified under **Newspaper Dailies** were specific to major dailies in Kenya. *#Natiomedia, #thestarkenya* are some examples in this category. **News Emerging issues** characterised breaking news, that could still be related to topics such as sports, politics, etc. As a result, 22 topics were manually identified corresponding to the list of hashtags in the tested data. Overall, this profiling process was meant to generate very specific interest profiles for users. For example, some users were interested in emerging news and sports and not politics. Thus, categorizing them under news may not have been very specific to their taste profile.

A subset of 2000 out of 5320 randomised tweets from the two categories were selected for labelling and subjected to the five neural network-based models for classification. Adjustable parameters were replicated in the classification algorithms for consistency. *Support Vector Machines (SVM), Boosting Tree, Decision Tree* and *Logistic Regression* algorithms were selected for the classification purpose. The idea was to select the best performing modelling framework whose embeddings could be subjected to Expectation Maximisation (EM) to generate soft topic assignments.

From the results in Figure 5.2, *Word2Vec-SkipGram* with 100 dimensions outperformed the other algorithms after the merger of the overlapped topics. Therefore, *Word2Vec-SkipGram (100 dimensions)* was selected as the modelling technique of

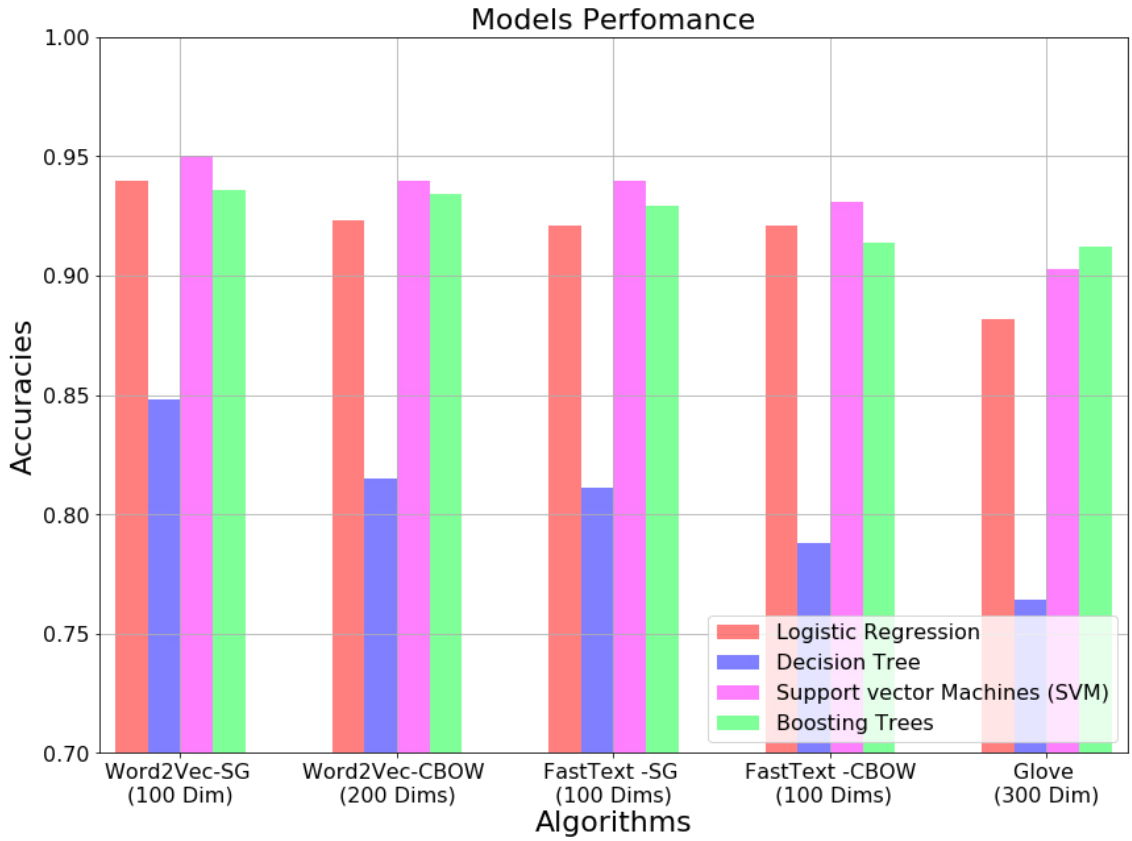


Figure 5.2: Model Classification Results

choice for further EM experiments where the *FastText* based ones were close in performance. This performance on the selected test data could have been influenced by other factors including the number of test tweets as ,normally, the larger the dataset, the better the resultant vectors. Therefore, *Word2Vec-SkipGram (100 dimensions)* performed better on the 2000 tweets. This was in relation to the proof of multi-interest modelling concept. Technically, soft cluster assignments from the rest of the algorithms was not very meaningful at this stage as *Word2Vec-SkipGram (100 dimensions)* generated the best set of embeddings for soft cluster assignments in this setup.

5.5.2 Cluster Numbers Heuristic Results

The *Word2Vec-SkipGram* model in Section 5.4.2 was trained with the following parameter specifications. ***Size=100,min_count=3*** and ***window=10***. Default *Word2Vec* values were assumed for the rest of the parameters. The output of the modelling process was a vectorised dictionary of 140252 words.

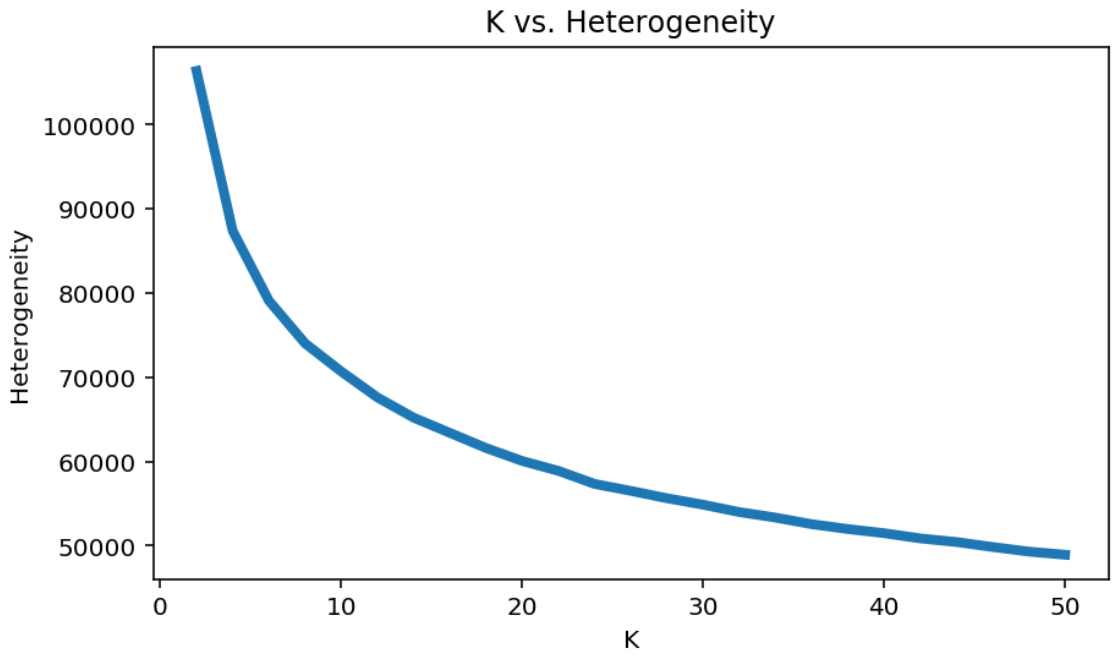


Figure 5.3: Elbow Heuristic Results

The optimal number of clusters in tweets for EM were computed following the Elbow heuristic (Bholowalia & Kumar, 2014) measure. To compute this, several K values representing probable number of clusters were factored. For each value of K, *K-means++* was applied to calculate *heterogeneity*. Heterogeneity is a measure of compactness in the clustering process. Ideally, the optimal cluster numbers should be as close as possible to the number of topics identified in Section 5.4.2. Results shown in Figure 5.3 depicts K to be a value between 20 an 30 corresponding to the 22 topics representative of the dataset as derived in Section 5.5.1 as the representative topical numbers. However, overlapping cluster centroids for 12 topical classifications were

merged with the rest of topics, as they were contextually similar by manual observation, leaving ten topics. Details of the merging steps are highlighted in Chapter 3. The resultant centroids were then extracted and used as initial means for EM.

5.5.3 Methodology Validation

To further validate the methodology and results of the EM process, a validation dataset of 282 users drawn from the Kenyan twitter space was used. 82 of the users are current politicians. Thus, the assumption is that their interests are largely political. It is worth acknowledging that politicians do not just disseminate political content. They also often tweet about current affairs in the country, development projects they undertake, etc. However, with EM, the goal was in quantifying the interest in specific topics. Therefore, if the interest that a user labelled as politician has in the **politics** topical cluster is higher/close to the highest classification value, then it is enough proof that the user is a politician. However, this may not be the same across all the users. Output of this validation process is a responsibility matrix depicting soft cluster assignments for the users in question. For each user, average interest vectors per topic were computed representing the user's profile as described in Section 5.3.4. For example, user *kithurekindiki*, a senator, tweeted only four times during the data collection period. Averaging the user's interest in the extracted topics, the user's interest was 100% in *Political Discourse*. On the other hand, a user such as *mutuamuluvi*, though with very few tweets, the diversity in topics was evident with 42.9% of his tweets being profiled as *Political Discourse*, 7.7% in *Life and Well-being*, and 8.6% *Condolences* among others. A subset of the entire output is shown in the Table 5.1. Usernames in the table have been anonymised for privacy reasons. The same output in Figure 5.4 shows that on average, **67.259955%** of content disseminated by the politicians is political. This affirms that the framework largely classifies correctly.

Usernames	Life and Well-being	Development Projects	Sports	Counties	Swahili Chatter	Newspaper Dailies	Political Discourse	Elections	News & Emerging Issues	Condolences
xxxnchxxxixxx	14.650000	0.000000	0.000000	0.000000	27.344231	1.865385	49.434615	1.911538	0.015385	4.776923
xxxxs_xxxi	2.732292	0.000000	0.000000	0.000000	3.599479	0.203646	92.355729	0.000000	0.000000	1.109896
hxxxenxxx	8.312000	0.000000	0.000000	0.335000	14.804000	1.003000	72.578000	0.000000	0.000000	2.968000
xxxengmuxxx	12.070186	0.000000	0.588199	0.030435	28.661491	2.173913	53.783851	0.439130	0.168944	2.080745
xxxkxxxdo	21.471809	0.000000	0.108511	0.000000	0.063298	3.369149	57.862766	5.003191	0.946809	11.169681
xxxuingixxx	18.790526	0.000000	0.420526	0.000000	4.048421	5.322105	64.768947	0.000526	0.557895	6.089474
xxxojaxxxon	5.366327	0.000000	0.000000	0.000000	13.888265	1.035714	78.695408	0.000000	0.000000	1.011224
xxxte_kexxx	9.029592	0.510204	0.497959	0.080102	10.478061	1.841837	73.866837	0.000000	0.000000	3.692857
xxxcentxxxnga	18.246875	0.520833	0.000521	0.000000	11.870833	0.958333	63.434375	0.001042	0.000000	4.965625
mpxxxkahaxxx	18.752717	0.000000	0.000000	0.000000	2.834783	6.498913	65.819022	0.525543	0.002174	5.561413

Table 5.1: EM Topical Classifications

Results in such a modelling process are applicable in follower - followee recommendations, dealing with cold-start recommendations and third-party content propagation in short-text microblogs.

5.5.4 Human Validation

To validate the results obtained in Section 5.5, 3 human evaluators were given the task. The evaluators were selected based on their knowledge of the English and Swahili languages, i.e., in deciphering content and assignment to a probable topic as well as their familiarity with specific domains, e.g., sports betting. General knowledge about Kenya was also a requirement, as some tweets required someone well versed with Kenya to discern their true contextual representation. They were presented with 20 tweets from 10 random tweeters, in the test set for validation. The evaluators were also presented with a curated list of topics in the test set outputs as in Table 5.1. They were expected to classify each tweet based on the presented subset of the identified topics. For consistency purposes, the extent, i.e., percentage of influence the topic had on the validation tweet, was not considered. The evaluators' job was simply to identify whether a tweet was relevant to the presented topics in line with the soft clustering approach. The results were presented the same way as in Chapter 4 where $X1$ to $X3$ represented the evaluators. x_{1u} to x_{3u} were the individual soft topic classifications as per the evaluators. $M_{1u}..M_{3u}$ represented the model's topical assignments for the

same tweet. The Kappa score was then computed representing the agreement between classifications in the model and human evaluators (McHugh, 2012).

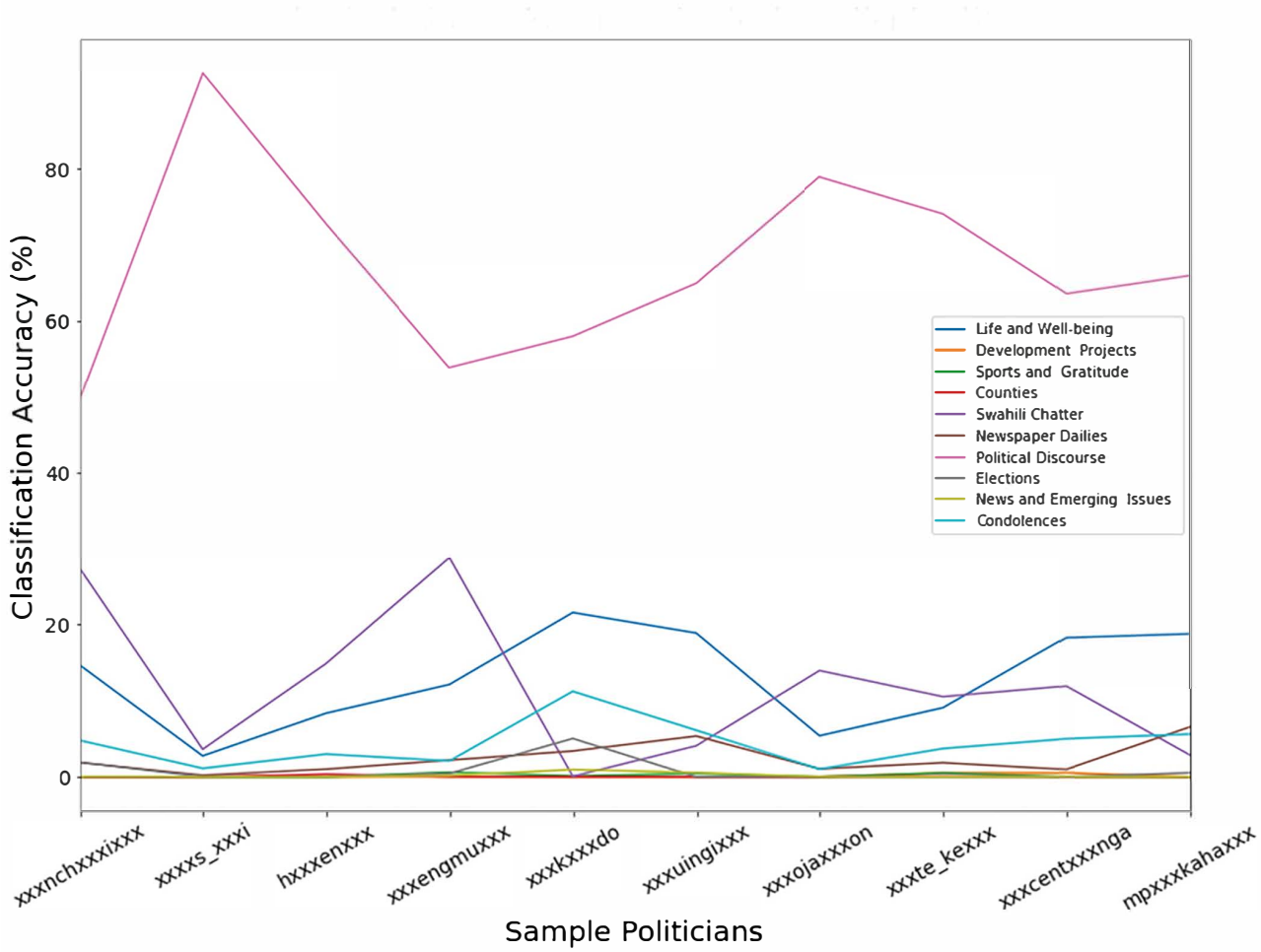


Figure 5.4: Politicians EM Classifications and Relevance to Political Content

k_{1u} to k_{3u} represented the Kappa score across the topics. Therefore, Kappa score k was computed as $k = (p_o - p_e)/(1 - p_e)$ where p_e is the hypothetical probability of chance agreement. On the other hand, p_o is the relative observed agreement between the model's and evaluators' observations.

Table 5.2 shows the classification agreements between evaluators and the model in the approach. From the table, the agreement between the model's and evaluators'

	X_1			X_2			X_3		
	X_{1u}	M_{1u}	K_{1u}	X_{2u}	M_{2u}	K_{2u}	X_{3u}	M_{3u}	K_{3u}
Life & Well-being	4/20	3/20	**0.828**	3/20	3/20	**0.608**	4/20	3/20	**0.483**
Development Projects	2/20	1/20	**0.643**	1/20	1/20	**1.0**	1/20	1/20	**1.0**
Sports	6/20	6/20	**0.762**	6/20	6/20	**0.524**	6/20	6/20	**0.524**
Counties	3/20	4/20	**0.483**	4/20	4/20	**0.688**	4/20	4/20	**0.688**
Swahili Chatter	1/20	1/20	**1.0**	1/20	1/20	**1.0**	0/20	1/20	**0.0**
Newspaper Dailies	5/20	6/20	**0.625**	5/20	6/20	**0.625**	5/20	6/20	**0.875**
Political Discourse	14/20	12/20	**0.783**	13/20	12/20	**0.681**	10/20	12/20	**0.4**
Elections	0/20	2/20	**0.0**	2/20	2/20	**0.44**	2/20	2/20	**0.44**
News & Emerging Issues	9/20	7/20	**0.588**	6/20	7/20	**0.659**	5/20	7/20	**0.529**
Condolences	0/20	1/20	**0.0**	0/20	1/20	**0.0**	0/20	1/20	**0.0**

Table 5.2: Kappa Scores *K* depicting rating agreement between the model and human evaluators

classifications was 59.2%, depicting a *moderate* to *substantial* agreement in the classifications as per the Kappa statistic scale (Viera, Garrett et al., 2005). This is a very strong indicator of profile consistency because only 20 tweets were considered for each user. A larger number of tweets would have likely resulted in a better Kappa score.

5.6 Chapter Summary

Definitive profiling of user's long- and short-term interests based on their disseminated content is possible in short-text micro-blogging platforms. The propagated content to a greater or lesser extent depicts certain topical preferences over time. The proposed framework optimises vector representations in classifying tweets as soft clusters via Expectation Maximisation, thus identifying the user representative interests for each user. For computation of a user's multi-interest profile, individual topical classifications

were aggregated. Validation of the framework's classifications was done via assignment of random tweets belonging to a few tweeters for manual topical classifications with the help of evaluators. Furthermore, a validation of a group of users with known perceived interests was done in the case of *political content*. The validation depicted a strong agreement (**0.592** Kappa Score) between the model and evaluator's classification.

This modelling process provides an enhanced model for user interest profiling, and and sheds light on the design knowledge of such models. For instance, it was demonstrated that the use of vector representations to learn short and often misspelled words in the disseminated content will potentially enhance the effectiveness of user interest profiling. Variations of user interests and with a determined certainty are generated. The framework in Chapter 4 aims at quantifying the degree of interest that users on short-text microblogs have towards certain topics of interest. However, the framework in this chapter went further in the depiction of user interests across several topics of interest as probabilities. With this approach, it is possible to discern a short-text microblog user's interest across several domains for better profiling. Ideally, it is possible to quantify these levels of interest across topics, for better third-party content or friendship network recommendations.

Time is an important factor in short-text microblogs, based on the variations in topical decay and gain on the platforms. Therefore, incorporation of timestamps in extraction of user interests is vital. This is because user A's representative interests today, may be different in a few days' or weeks' time. Therefore, topical volatility on such platforms is high. In Chapter 6, the element of semantic changes modelling is introduced with time as a factor. This results in not only accurate, but time-sensitive user-representative profiles.

Chapter 6

Multi Interest Semantic Changes

6.1 Introduction

Consumption of content in short-text microblogs is necessitated to a large extent by individual users and their friendship network interests. Based on the dynamism in the data throughput on such platforms, e.g., Twitter, prevailing conditions are bound to determine the nature of consumed or disseminated content. Therefore, semantic interests differ over time even for individual users. Detecting this semantic change over time is integral in mapping user profiles over a time frame, especially in microblogs where only the extrinsic user profile identifiers provide metadata that seldom evolves.

Currently, content propagation has been proliferated by the surge in citizen journalism. This is partly attributed to the increase in the number of devices e.g. mobile phones, access to internet as well as the emergence of many social microblogging platforms like Twitter. Generally, tweeters[1] consume content on the platform based on their prevailing interests at the time. For example, in times of political campaigns, many demographically relevant tweeters are likely to express interest in political content. However, this interest is likely to decay over time when the political season is over.

[1] *user who posts tweets on the Twitter online messaging service*

The same scenario could be replicated in a sports season where support of teams fades as the season winds up. Modelling and extraction of such user dissemination patterns and representative interests is a challenging task especially for legacy systems. This is attributed to two factors *(i) The data volatility factoring its dissemination throughput (ii) Time-based variations in the nature of topics of interest.*

The ability to extract and present time-sensitive and accurate user representative profiles from such evolving short texts is important in recommender systems design research. The design goal of recommender systems on such platforms is to personalize both the third-party content, and the user identification process. This in turn presents the most relevant users as follower-followee suggestions, as well as delivery of more personalized third-party content for the users. This personalisation process is based on the extrinsic interests from the disseminated content.

With the personalizing goal, there is need to develop a framework that is able to capture user-representative interests on such platforms, but with a factor of time. Such a framework has the ability to present semantic profiles to third-party content curators from an informative point of view on how user interests evolved with time. This can serve specific purposes related to the generation of profile information for users of interest in certain topics at given times. Such interest patterns can then be used in recommendations of time specific content, as well as related follower-followee networks. In the development of this framework, the following research questions are addressed :-

- Is it possible to extract user-representative interests in time-series based streaming short texts for profiling?
- Are the extracted patterns in the short-text sufficient in making time-dependent user/content recommendations/predictions for generalized short-text content disseminators?

6.1.1 Notations and Symbols

The below notations and respective descriptions are used in this chapter: -

θ_x - A topic proportion's vector for each document x.

σ_v - Multi-dimensional vector representation for term v.

v - a term in a corpus.

t_x - Variational timestamp of a document/tweet x.

$\mathcal{N}$ - Logistic normal distribution over (O, I).

(O, I) - Gaussian variables.

6.2 Problem Statement

The assumption in the design of the framework in this chapter is that, the semantics of disseminated content change over time for individual users. Therefore, a framework with the ability to discern semantic user interests of short-text microblog users is presented [2]. This is challenging in short-text microblogs as the level of expressivity is not always exhaustive due to character limitations per document on such platforms. This is in addition to factors related to throughput and content decay and gain over time on such platforms. For example, the text attribute in a tweet's metadata is limited to 280 characters, though realistically, the average tweet length is much shorter.

In the quest to design such a framework, there was a need to first extract topical representations of the data at specific timestamps. Since documents were short-text, vector representations worked well in discerning their semantic relevance. Each word token in the tweet was modelled as the inner product between word and topic embeddings as vectors, across specific periods. Topics of interest at each period were captured with the overall semantic representation being the user interest weights across the dataset collection period. This way, semantic divergence across several topics of interest were

[2] Submitted to the Knowledge-Based Systems Journal

accurately represented.

This work is inspired by previous research in the computation of the Degree of Interest in Sports Betting (DoiSB) in Chapter 4 as well as multi-interest User Profiling in short-text microblogs in Chapter 5. In follow-back recommendations, user-representative interests recommendations were made among short-text microblog users with shared interests. The social *theory of homophily* was applied in validating the semantic correlations among the users. In multi-interest profiling, the quantification of user interests across topics was computed by generating a *responsibility matrix* across users depicting their interest levels across the topics. Algorithmically, Expectation Maximisation (EM) and Gaussian Mixed Models (GMM) were applied over the vector representations to extract soft clusters. Furthermore, the semantic distance to the clusters per user-aggregated vector representations was assumed to be the interest level in the topical cluster.

However, this work differs from the above as it incorporates timestamp-based topic and word embeddings on short and evolving texts in identification of evolving user interests.

The below points highlight the novel aspects of research in this chapter: -

1. The combination of word and topical embeddings as a time-variational distributional model differs from other works in the user profiling domain. Conventional modelling in related works in this domain, point to either word or topical representation models and not an amalgamation of the two.

2. In this approach, time sensitivity is definitive in the representation of the generated user-representative profiles. Inferencing topics over word embeddings where topical vectors are generated per mini-batch of documents and timestamp, highlighting the dynamicity in identification of representative user interests in short texts. This time-sensitive approach differs from keyword, concept and hybrid

profiling approaches that make use of external knowledge-bases in extraction of such interests. Vocabulary sparsity, volume, variety and velocity in short texts dissemination makes the above profiling approaches insufficient in this scenario.

In addition to the above, the following contributions were made in this chapter : -

1. Formulation of semantically representative user profiling framework that considers the disseminated content as time-based entities.

2. Each word in the framework is modelled as a categorical distribution of word embeddings and a time-based representation of the word's assigned topic.

3. The model and data ingestion framework is tested on a generic set of tweets geolocated to an area over time. This ensured accuracy in validation of the end results using a true class dataset over certain periods in the collection. User-representative interests over a generic test set were computed by the methodology that qualitatively outperformed the other approaches across a range of measures.

4. Quantitatively, semantic weights between the control and test sets in five sub-topics across ten quarters were measured depicting their semantic correlations. Linearity in the correlations across the timestamps indicated the validity of our modelling approach.

This work and related output is pertinent especially for third-party content propagators in the following aspects: -

- Dissemination of semantically relevant time-variational content to users on the short-text platforms.

- Accuracy in the forecasts of the type/nature of content users are bound to consume in certain demographics and the likelihood of their future consumption by learning their current consumption patterns.

- Content engagement patterns over time as content related to certain topics may be of more interest at specific times.
- Identification of the most relevant users to serve content, e.g., from third-party content disseminators. This is relevant in cold-start scenarios too.

6.3 Multi-Interest Semantic Changes Framework and Summary of Literature

In modelling evolving user interests, the proposed framework encompasses several processes related to the generation of *topical interests, word embeddings* and a *time-variant distributional model*. The topical interests are distributed over word representations as time variational topics in the model. Time variations are user-defined as the dataset is time-series documents. Latent Dirichlet Allocation (LDA) is used to generate the topical information at each timestamp (Blei et al., 2003a). Summaries of the most optimal K topics are generated through a discrete probability distribution over terms. The topic model output is vectorised in the time-based distributional model as described in Section 6.3.2. This mitigated the vocabulary sparsity problem with LDA (Tajbakhsh & Bagherzadeh, 2019).

The core processes in the proposed framework are discussed below and the interrelations between the processes captured in Figure 6.1 :-

1. **Dataset Collection** - A generic set of tweets were collected via Twitter's search API [3] for pre-processing and further modelling.
2. **Short-Text Modelling** - Inputs to the modelling module were word tokens from

[3] https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets

the pre-processed tweets. Timestamp information was extracted from the tokenised tweet. Neural-network representations via *FastText* (Bojanowski et al., 2017), *Word2Vec* (Mikolov, Sutskever et al., 2013) and *Glove* (Pennington et al., 2014) were made use of in vectorising the tokenised tweets. For comparative purposes, LDA (Blei et al., 2003a) and Twitter-LDA (W. X. Zhao, Jiang, Weng et al., 2011) were also used in the computation of topical qualities as baselines. Other variants such as Latent Semantic Analysis (LSA) (Landauer, Foltz & Laham, 1998) and lda2vec (Moody, 2016) exist in the topic modelling domain. However, comparative inaccuracies between LSA and LDA for example, made LDA a preference. On the other hand, lda2vec uses Word2Vec in its vectorization. However, Word2Vec models words atomically which makes it insufficient in tweets that are often mispelled. There were variations in input parameters to these models as described in Section 6.4.1.

3. **Topics over Word Embeddings Inferencing** - To model topics over embeddings for each term, a distribution model of topical representative words over word embeddings was first computed. The output was the inner product of topic and word embeddings at each period. This provided for a better generalisation of the topics. This process allowed for smooth variations of topics over the specified period in the dataset.

4. **Multi-interest Semantic Changes** - At each timestamp, word weight probabilities were computed. The probabilities represented the word semantic weights over the vectorised set topics. Therefore, the generated topics at each timestamp were assumed to be topics of interest at that specific time. The interest changes were captured by weighting interest keywords across the timestamps. Positivity in the semantic change indicated topical gain while the reverse represented topical decay.

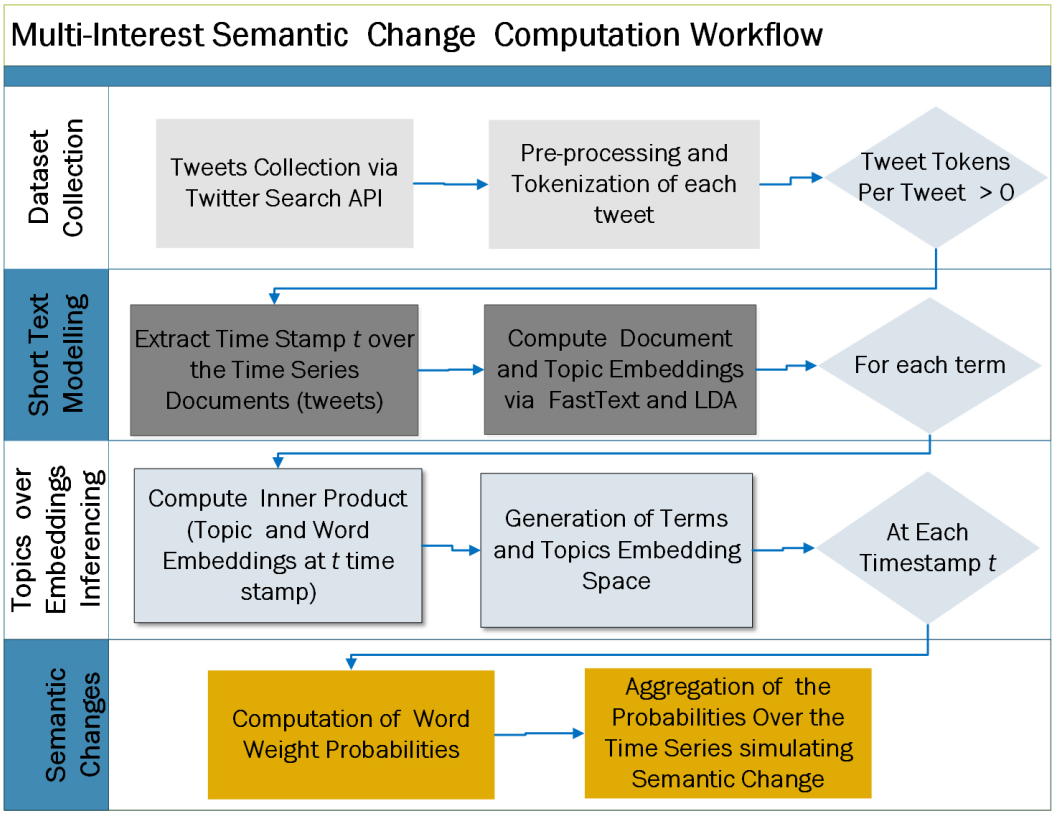


Figure 6.1: Multi-Interest semantic changes computation framework.

Figure 6.1 summarises the multi-interest semantic changes detection process as detailed above. As depicted in the framework, topical diversity and capture of terms that are representative of those topics at different timestamps encompasses a few processes. Latent Dirichlet Allocation (LDA) is applied in computing the topical representations over vectors. Ideally, LDA on its own doesn't perform well on short texts as its modelling process is based on vocabulary co-occurrences, which is a shortcoming in short texts (Chang et al., 2009). Therefore, the vector representations in the distributional model, helped capture semantics better. A brief description of the LDA modelling process follows in Section 6.3.1.

6.3.1 Latent Dirichlet Allocation (LDA)

In the proposed approach, LDA is the core topic modelling approach in the distributional model at each timestamp. As mentioned in Section 3.2.1, it is an unsupervised technique in the discovery of knowledge in form of coherent topics in text where tweets in this instance are represented as a bag-of-words (Blei et al., 2003a). With LDA, a summary of pre-set topics is computed through a discrete probability distribution over words. A per-document distribution over the generated topics is then inferred. In this setup, each tweet is a document. Term co-occurrence likelihood is the basis of topic formulation in LDA. However, its performance is affected when the vocabulary is sparse just like in tweets that are often shortened (Chang et al., 2009; Tajbakhsh & Bagherzadeh, 2019). To put this in perspective, LDA considers k topics, each of which is a distribution over the vocabulary in the corpus. Furthermore, there is consideration of a topic proportion's vector θ_x for each tweet or document x in a corpus of X documents with V distinct terms. Let $b_{xn} \in 1, ..., V$ denote the n^{th} word in the x^{th} document. In the topic generation process, each term v is assigned to a topic k with a probability θ_{xk}.

Vector Representations:

Vector representations are integral in extraction of semantic knowledge from both short and conventional texts. In the selection of the best vector representation framework, a few neural language modelling approaches just like in Chapter 5 were adapted. Comparative experiments were then carried out to further select the neural language model of choice in the distributional model based on a few quality measures. The dataset was trained with *Word2Vec* (Mikolov, Sutskever et al., 2013), *Glove* (Pennington et al., 2014) and *FastText* (Bojanowski et al., 2016) based embeddings. A comparison with other conventional topical modelling algorithms was also carried out. The embedding algorithms were experimented as baselines with different but consistent dimensions.

Results on their topic quality measurements as in Table 6.1 informed the embedding framework of choice for further modelling.

FastText algorithm modelling unlike other variants ignores word structures formulation. With FastText, each word token w is represented as a bag-of-character *n-grams* in the vector space. The word token itself is also included in the n-grams (Wandabwa, Naeem, Mirza & Pears, 2020). To incorporate most of the n-grams, $3 \leqslant n \leqslant 6$ are optimal (Bojanowski et al., 2017). With an n-gram dictionary of size Y and word w, $Y_w \subset \{1, ..., Y\}$. x_y is the vector representation of each n-gram b as in the below equation. A word is represented by the sum of the vector representations of its n-grams. Thus, the scoring function is formulated as $s(w, b) = \sum_{y \subseteq Y_w} x_y^\top v_b$ where v is the word vector and b the context position of the word. Unlike *FastText*, *Word2Vec* and *Glove* modelling ignore word morphology i.e. the character sequence that forms a word, thus each term is treated as whole i.e. atomically. This makes vectorization of short and noisy text difficult in such modelling approaches.

6.3.2 Topics over Word Embeddings Inferencing

As mentioned in Section 6.3.1, advantages of word vector representations and topic models were leveraged in this phase. This was pertinent in the distribution of topics as user interests over time variations, accommodating datasets that span over periods of time. Vector representations of terms and topics were modelled in the distributional model as follows: -

1. For each term v, the model considers an M-dimensional vector representation σ_v. At each timestamp t, the distributional model embedding is presented as $\alpha_k^t \in \mathbb{R}^M$ for each topic k. This meant that each topic, with relevant probabilistic terms is denoted as a time-variational vector with documents separated in mini-batches (Dieng, Ruiz & Blei, 2020). Therefore, the probability of a word in

document x (as in Section 6.3.1) to be modelled in given topic, is computed as the exponential inner product between the topic and word embeddings at corresponding timestamps (user defined) in the dataset. This probability function is represented in Equation 6.1.

$$P(b_{xn} = v|z_{xn} = k, \alpha_k^{(t_x)}) \propto exp\{\sigma_v^\top \alpha_k^{(t_x)}\} \tag{6.1}$$

z_{xn} is the topic assignment in LDA for each word n in the document as described in the probabilistic extraction of topics in Section 6.3.1.

2. Correlation between terms and their representative topics is higher when the term and topic embedding is also higher in the embedding space. Terms with such similarities just as in clustering, are bound to be grouped together. Markov Chain (Besag, 2004) enforces this smooth variation of topics over their embeddings. In addition, time priors are considered over the vector of topic proportions α_x. The model then captures variations in topics over the time series dataset. The prior over the vectors of topic proportions depends on a latent variable λ_{t_x} where t_x is the timestamp variation of document/tweet x. This probabilistic process is represented as $P(\theta_x|\lambda_{t_x}) = \mathcal{N}(\lambda_{t_x}, Q^2 I)$ where $P(\lambda_t|x_{t-1}) = \mathcal{N}(\lambda_{t-1}, \sigma^2 I)$. Q is a model parameter, while σ is a hyper parameter controlling topical smoothing of the Markov chains. $\mathcal{N}$ is the logistic normal distribution that transforms Gaussian variables (O, I) to the simplex.

   Overall, the combinatory modelling approach where terms are distributed over word and topic representations differs from conventional topic models. The difference is in the probabilistic distribution of topics over the vocabulary in the corpus just like in LDA.

The modelling process is captured in Algorithm 5.

Algorithm 5 Embedded Interest Topic Modelling Process over Time

Input: Tweet Tokens v, Corpus X, Time Stamp $t_1, ..., t_x$

Initialisation of model parameters ▷ Dataset, model and model optimisation arguments (t = $1, ..., T$,k = $1, ..., K$)

1: **for** i = $Iterationcounts$ **do**
2: Compute document(tokenized tweet) and topic embedding means
3: Compute the topics ζ_k^t = $softmax(\sigma^\top \alpha^t{}_k)$
4: Extract tweets in mini-batches as per the set dataset argument
5: **for** each tweet x in each mini batch **do**
6: Compute the topic proportions θ_x
7: **for** each word c in each tweet **do**
8: Compute $P(v_{xc}|\theta_x)$ = $\sum_x \theta_{xk} \zeta_{k,v_{xc}}^{t_x}$
9: **end for**
10: **end for**
11: Update the dataset, model parameters and hyper-parameters
12: **end for**

In Algorithm 5, the model inputs are tweet tokens as time series documents. Modelling the tokens requires adjustment of model parameters, and hyper-parameters specifications (Line 1). Depending on the training iteration counts, terms and topical means are computed (Line 2,3 and 4). The word embeddings were generated as described in Section 6.3.1. For each list of tweet tokens representative of a tweet in each mini-batch, topic proportions are computed via LDA from the topic embeddings and the closest terms assigned to the most representative topics (Lines 6 ... 9). The closer the topic and term means are, the higher the likelihood of being clustered together. Therefore, semantically close terms are assigned to similar topics as their representations are close in the embedding space. The process is repeated for all terms across the mini-batches until all tokens are modelled and all iterations completed. Model and data parameters as well as hyper-parameters can then be updated and the modelling re-run until convergence.

6.4 Experimentation

This section validates the processes presented in Section 6.3. A few consecutive steps are followed in the modelling process in addition to the description of the time-variant dataset in the experimentation. The experimentation process aims to generate topics as interests that are sensitive to time spanning the dataset period. Therefore, the end result in this experimentation phase is a comparative evaluation of an agreement between the control set in Section 6.4.1 and the generated topics from the dataset over the test set period.

6.4.1 Datasets

The goal of the experimental process was to compute topical interest evolution over a dataset. This meant better extraction of user interest/preference changes over time in short texts. To simulate this process, a dataset of tweets that bore the following characteristics was collected :-

1. **Generic set of tweets** - Tweets were collected via Twitter's search API. The geolocated but generic set of tweets was collected from Kenya. Ideally, this presented a set that could be validated via a control set as described in Section 6.4.1. This does not mean that the framework is specific to just one location. As long as the dataset is of short nature and timestamp-based, this approach is applicable.

2. **Language independence** - Approximately, 90% of the tweets were disseminated in English. The rest were in Swahili and a mixture of the two. The modelling process especially with embeddings was language independent as context was applied as opposed to vocabulary co-occurrences.

3. **Time Variance** - Extracted tweets spanned a period of five years segmented in

quarters, i.e., from *2015 Quarter One (Q1)* to *2020 Quarter Two (Q2)*. Timestamp variation t_x for each tweet was incorporated.

Training Set

The models were trained on a corpus of 828,789 generic, cleaned, and tokenised tweets with the collection geolocated to Kenya as mentioned in Section 6.4.1. The choice of Kenya was influenced by the availability of a control set. The same collection can be replicated across different geographical regions for extraction of larger datasets especially if the control set is not of essence, unlike in this scenario. The collection period was between *2015 Quarter One (Q1)* and *2020 Quarter Two (Q2)*. Five quarters were excluded from the dataset as very few tweets were collected in the respective quarters. These were **Quarter 4 of 2015, Quarters 2 and 4 in 2016 and, Quarters 3 and 4 in 2017**. Less data with short very sparse vocabulary made it impossible to keep the data in the collection. Therefore, the resultant dataset spanned 17 quarters across 2015 to 2020. Each tweet in the collection included the tweet's metadata, e.g., geo-coordinates, hashtags, retweets, etc.

Control Set

A *control set* was needed to validate the conversation in Section 6.4.1. This was to provide a semantic proof of topical evolvement over time (Hassan, Arslan, Li & Tremayne, 2017), (Hannak, Margolin, Keegan & Weber, 2014). The assumption is that the semantics in news items positively correlate with the disseminated tweets in specific demographics (Sankaranarayanan, Samet, Teitler, Lieberman & Sperling, 2009; Brena et al., 2019). Furthermore, semantics in tweets and to a large extent evolves with changes in the news items.

Since the initial collection was geolocated in Kenya, it was prudent to collect news items in the same demographics as a control set. Therefore, $189,906$ tweets

from Twitter handles of two major media houses, i.e., *Nation Media*[4] and *The Star-Kenya*[5] were collected for validation. Retweets and tweets with just mentions of these Twitter handles were not incorporated in the collection. The aim was to collect tweets solely disseminated by the media houses as the assumption was that they semantically correlated with the initial generic collection.

Test Set

A neutral but relevant set of test tweets was needed in the testing phase of the framework. The goal was to make sure that the model worked to the expected level on a neutral dataset. A collection of $161,675$ generic tweets from the same geographical bounds as the training data was collected. The test set was not part of the training and control sets. In addition, each tweet entry also contained associated metadata such as hashtags, mentions, etc. Retweets were filtered out of the dataset as they made up the bulk of the replicated tweets.

6.4.2 Parameters and Model Training

The proposed framework is segmented into two important parts. The vector representation part and the topic modelling one. The two had diverse but consistent parameters across the different algorithms as used in the experimentation process. The parameters are detailed as in the sections below: -

Word Embedding Settings:

The neural language models were adopted in generating vector representations over the training set in Section 6.4.1. The idea was to trial out several modelling algorithms and pick the best performing one for further modelling. This meant subjecting the training

[4]https://twitter.com/dailynation
[5]https://twitter.com/thestarkenya

data to *FastText*, *Word2Vec* and *Glove* algorithms for vectorisation and comparisons in performance. *FastText* can extract syntactic information from a textual corpus regardless of the language of expression and misspellings. It does not ignore the word morphology, which is a limitation in short texts, with limitations in word co-occurrences. With *FastText*, vector representations are associated with each character n-gram. Words that are typically made up of characters are then represented as the sum of character vectors computed using a sliding window. This is one property that makes *FastText* work well with misspelled or shortened words such as tweets.*Word2Vec* and *Glove* modelling works the same way with a slight variation in the term vectorisation process. Words are modelled atomically and not at character level when compared to FastText. In the generation of embeddings, the *number of dimensions, learning rate, context window size, minimum count, window* and *epochs* were specified. They were as follows for *FastText* and *Word2Vec* models:

- *Size* of the dimensions. This ranged from 100 to 300 dimensions consistent with (Mikolov, Chen et al., 2013)
- *min_count* or minimum count of a word in the corpus. Any word with a co-occurrence count less than the specified parameter is not incorporated in the training set;
- *sg* parameter for training a skip-gram model if $sg = 1$, otherwise Continuous Bag of Words (CBOW).
- *word_ngrams* to enrich word vectors with sub-word (n-grams) information if specified as 1 ;
- *iter* or iterations. This specifies the number of epochs over the corpus.
- The *window* parameter specifies the maximum distance between the current and predicted word in a tweet;

- *Glove* model only provisions for the *epochs* and *learning rate (lr)* to be defined.

Topic Modelling Settings:

As stated in Section 6.3.2, the end result was a distributional model that encompasses modelling of topics over embeddings, with time as a factor. Variances of different priors $\lambda^2 = \sigma^2 = 0.005$ and $Q^2 = 0.5$, consistent with the baseline setup in (Blei & Lafferty, 2006; Dieng et al., 2019a) were adapted. The optimal number of topics was 10 after 50 passes across the models, including in LDA and Twitter-LDA baselines. This was based on the Elbow heuristic (Bholowalia & Kumar, 2014) measure. To compute this, several K values representing probable number of clusters were factored in the modelling process. For each value of K, *K-means++* was applied to calculate *heterogeneity*. Heterogeneity is a measure of compactness in the clustering process. Parameters such as *batch size* = 50, *dropout rate* = 0.1, *learning rate* = 0.005 were sufficient in this setup after several iterations of different batch values. A fully connected feed-forward inference network for topic proportions α_x with ReLU activation (Y. Li & Yuan, 2017) was used. The training was run over 20 epochs.

Model Generation

As mentioned in Section 6.4.1, the model was trained on 828789 tweets. The goal of the modelling process is to make it possible for words in the training to have contextual inference. This is important in the computation of inter-word/sentence distances. Ideally, words with close contextual similarity are likely to have a high co-occurence in the training set. For example, *"Donald Trump"* is likely to have a high similarity score compared to *"Donald Biden"* in the US political dataset.

Pre-processing the tweets as input corpus to the model followed these steps :-

- Lower-cased all terms in each tweet.

- Filtered out numbers and encoding accented characters. Numbers were not of importance in this modelling process.
- Hyperlinks in the corpus were removed.
- Removal of hashtags in the dataset. They are user-generated words representative of a topic of interest and are normally prefixed by the hash (#) symbol.
- User mentions were also removed. They are usernames on the platform prefixed by the @ symbol.
- Words with a character count less than three were removed. Their contextual significance was not high as they were made of prepositions, etc., that were not in the stopword list.
- NLTK stopword list [6] was used to filter out words in the list out of the dataset.
- Tokenisation of tweets where individual terms in each tweet are split and appended to a list for further modelling.

The output, which is a clean and tokenised set of terms in a list, was trained via the modelling algorithms mentioned in Section 6.4.2.

As depicted in Algorithm 5, the tokenised list of terms topical proportions was computed via LDA. This process was computed over each mini batch, which was ideally on a quarterly basis. The closest terms in each mini batch, were assigned to the closest topic. The semantic weight of a term in the LDA generational model relative to the model was extracted at each mini batch. The same process was repeated across the modelling algorithms and their variants over all mini batches. For consistency purposes, model generation hyper-parameters were all the same as in Section 6.4.2.

[6] http://www.nltk.org/

6.5 Results

The framework's performance was measured quantitatively and qualitatively to ascertain that the results corroborated in both dimensions. Quantitatively, topical quality across the timestamps was the tested measure using the generated embeddings. Ideally, the best performing modelling algorithm was adapted in further qualitative evaluations. This related to intra-topical changes over time and correlations in topical interests, a key measure in the user profiling process in streaming texts.

6.5.1 Topic Quality Evaluation

Two metrics were used to compute the quality of generated topics over time. **Topic coherence (TC)** across the topics in each specified period (Röder, Both & Hinneburg, 2015; Korenčić, Ristov & Šnajder, 2018) was one metric. Normally, topics are represented as sets of important words with semantic relevance towards a specific theme. Depending on the type of dataset, modelling algorithm and fine-tuned parameters, the output will either be topics that either make sense or not, as they need to be interpretable if they are to be judged by humans. Therefore, topic coherence measure distinguishes good topics from bad ones by computing the degree of similarity between high scoring words in the topics. Secondly, **topic diversity (TD)** was computed. Basically, this measures the percentage of unique words in the top 25 words of all topics (Dieng, Ruiz & Blei, 2019b; Dieng et al., 2020). A product of the two scores was an indicator of **topic quality (TQ)**. Results are captured in Table 6.1. Scores close to zero indicate redundancies in the generated topics, and thus the modelling algorithm is deemed inferior. Overall, $FastText$ based models performed better than $Word2Vec$ and $Glove$ variants in the embedding based techniques. This is largely attributed to their character-level modelling capabilities making them ideal for datasets with mispelled or shortened words reminiscent of tweets. However, FastText-Skip Gram with 100 dimensions performed

Model	*Size*	*TD*	*TC*	*TQ*
FastText Skip-Gram	100	**0.88**	0.7821	**0.6882**
Word2vec Skip-Gram	100	0.8	0.682	0.5456
FastText-CBOW	100	0.81	**0.847**	0.6861
Word2Vec-CBOW	200	0.86	0.73	0.6278
Glove	300	0.865	0.701	0.6064
LDA Baseline	10 topics	0.8	0.3833	0.3066
Twitter-LDA Baseline	10 topics	0.83	0.5545	0.4602

Table 6.1: Topic Diversity (TD), Topic Coherence (TC) and Topic Quality (TQ) values across five vector representation techniques and two topic modelling baselines (Dieng et al., 2019b)

best across the two quality variables. LDA (Blei et al., 2003a) as well as Twitter-LDA (W. X. Zhao, Jiang, Weng et al., 2011) baselines performed dismally in modelling of interpretable topics of interest. Therefore, *FastText-Skip Gram* with 100 dimensions was selected for further modelling of intra-topical changes over time as well as topical interests correlations.

6.5.2 Qualitative Evaluation

Qualitatively, two approaches were used in discerning changes in interest levels for users over time: -

1. **Intra-topical changes over time** - This process involved tracking of interest changes over time. Topical interests reflect the nature of the consumed content over a time period.

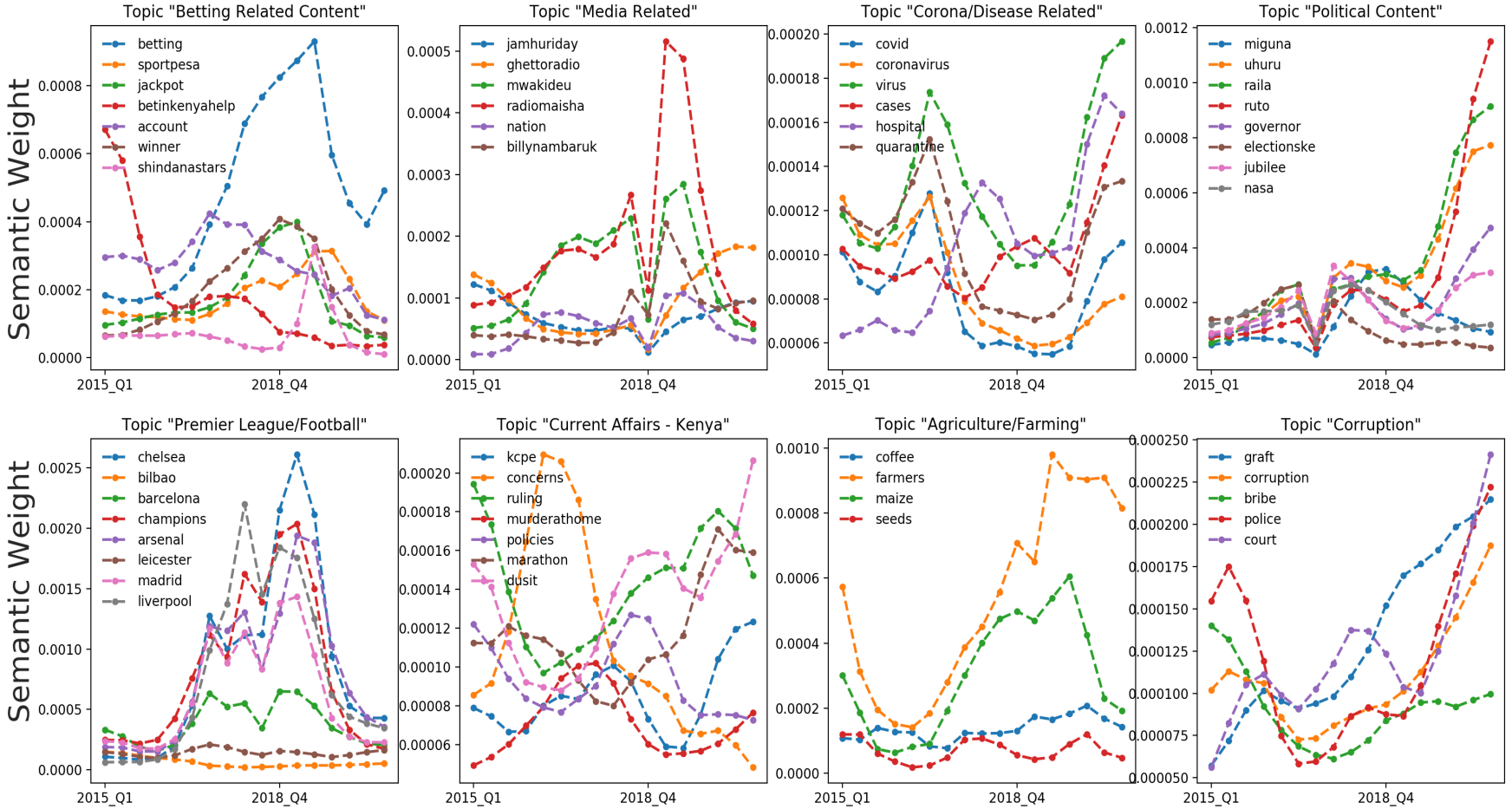
2. **Topical Interests Correlation** - The assumption is that topical interest weights in the test and control sets should positively correlate (Sankaranarayanan et al., 2009; Brena et al., 2019). Therefore, a correlation measurement score was computed to ascertain the correlations.

Intra-topical changes over time

A representation of the topical interests over time is depicted in Figure 6.2 from *2015 Q1* to *2020 Q2*, a period of 17 quarters as described in Section 6.4.1. The interest changes are depicted by the probabilities of specific topical interest words across the timestamps in the test set. Basically, the probabilities are simply weights/influence a word has in a topic and across the time periods. The same is the evaluation measure in the control set too. From the figure, it is evident that overall interest in, for example, *"Football"* by the audience in Kenya diminished in 2020. This is characterised by the lack of any footballing activity at this time when most European countries banned all forms of sporting activity [7], a preferred activity for the Kenyan population. On the contrary, sub-topics with terms related to *"virus","coronavirus"* and *"quarantine"* gained attention at the start of 2020. The weights of these terms are probabilistically relative to the topic across the dataset timestamps. This period was characterised by lots of chatter about the virus across the world, Kenya included. In addition, this chatter time coincided with the COVID-19 mitigation measures by the government of Kenya [8].

[7] https://bbc.in/2Onmvi0
[8] https://bit.ly/3ew5InE

Figure 6.2: Semantic Weights (Word probabilities) in eight topics across 17 quarters, i.e., 2015 - 2020 in the test data. Probabilities shift with variations in time representing the overall interest levels across time. Interest in a word like "betting" rose exponentially from *2016 Q1* until about *2019 Q2* in the Betting related topic. Scaling varied per graph for better representation of individual semantic weights variations as the values differed largely across individual graphs.



Topical Interests Correlation

As much as eyeballing gives an idea of the patterns in the dataset, it is not a scientific measure to draw any plausible statistical conclusions. To ascertain the topical interest change was a true representation especially over time, a control set was made use of as described in Section 6.4.1. This dataset was a collection of news items geolocated to Kenya just like the test dataset. The assumption is that news items are largely a reflection of what most tweeters disseminate (Sankaranarayanan et al., 2009; Brena et al., 2019). Therefore, agreement in the topical relevance in both the control and test sets is a possibility.

Figures B.1, B.2, B.3, B.4, B.5, B.6, B.7, B.8, B.9 and B.10 show the topical distributions in the control set. In this set, the topical changes span a period of ten quarters. The correlations between the control (news items dataset) and the test set was in the semantic weights meaning the topical interests across the timestamps varied. To ascertain the correlation between the control and test set weights, a correlation measure was computed at each timestamp between the two sets. For consistency, the computation was done for the last 10 quarters, a period the control set was available. Table 6.2 has records of these semantic weights in the two sets across the quarters. For demonstration purposes, five sub-topics were selected from the control and test sets as in depicted in Table 6.2. The choice was driven by the presence of the sub-topics across the dataset collection period.

Figure 6.3 is the resultant scatter plot of the subtopic weights in the test and control sets from *2018 Q1* to *2020 Q2*. From the figure, the values between the two sets largely correlated more so, in sub-topics such as "Diseases Related - Coronavirus" and "Diseases Related - Quarantine" as they are in the same semantic space in the plot. Their weights in both sets were closed to 0 in the initial 8 quarters. Since the weights are probabilistic relative to each topic, the assumption is that weight will still be slightly

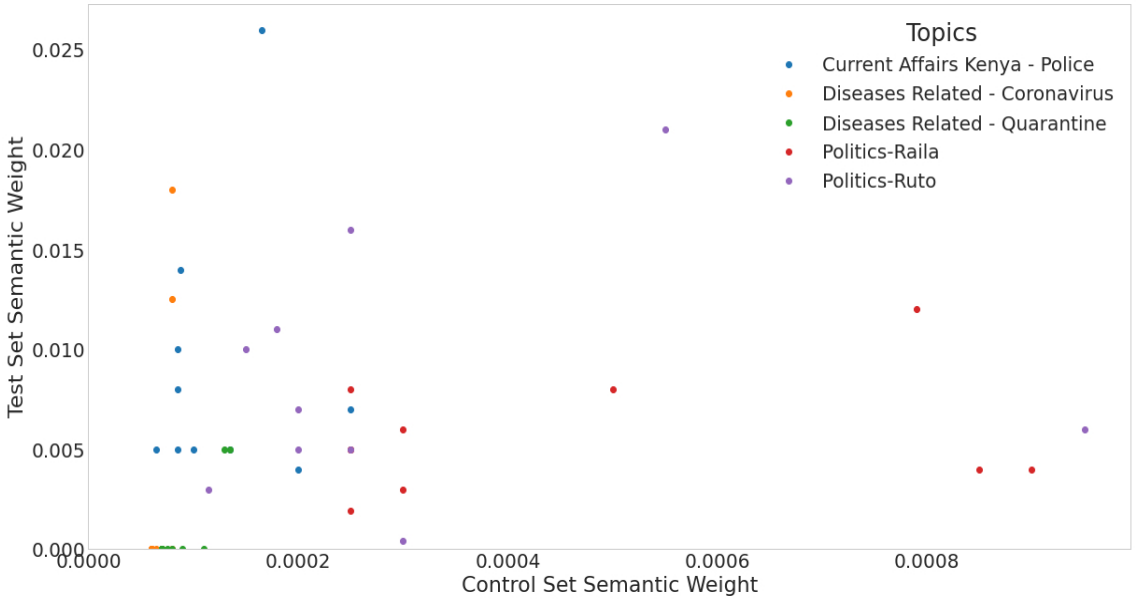


Figure 6.3: Semantic weights (Word probabilities) as the interest score in five subtopics across 5 topics between *2018 Q1* and *2020 Q2*. Each data point is an interest score for each of the subtopics in the test and control sets per quarter. Individual values per quarter are in Table 6.2.

greater than 0. This is attributed to the fact that the term at least exists in the topic, and thus, has an influence. This explains why their topical influence was very low prior to *2020*. Their semantic weights increase in the last two quarters, i.e., *quarters one* and *two* of *2020* the period in which *COVID-19* was declared an epidemic in Kenya. The same pattern in the subtopics is replicated in the test set as shown in Figure 6.2. On the other hand, *"Politics-Ruto"*,*"Politics-Raila"* and *"Current Affairs - Kenya"* depicted sinusoidal patterns over the 10 quarters in the two sets. This correlated to the nature of political content shared in the country over the time frame as shown in Figure 6.2.

	Sub-topic	Dataset	2018_Q1	2018_Q2	2018_Q3	2018_Q4	2019_Q1	2019_Q2	2019_Q3	2019_Q4	2020_Q1	2020_Q2	PCC
1	Diseases Related - Quarantine	*Test Set*	0.00009	0.000075	0.00007	0.00007	0.00007	0.00007	0.00008	0.00011	0.00013	0.000135	**0.871**
		Control Set	0	0	0	0	0	0	0	0	0.005	0.005	
2	Diseases Related - Coronavirus	*Test Set*	0.00008	0.00007	0.000065	0.00006	0.00006	0.00006	0.00006	0.00007	0.00008	0.00008	**0.635**
		Control Set	0	0	0	0	0	0	0	0	0.0125	0.018	
3	Current Affairs Kenya - Police	*Test Set*	0.000065	0.000085	0.000088	0.000085	0.000085	0.0001	0.000135	0.000165	0.0002	0.00025	**0.090**
		Control Set	0.005	0.01	0.014	0.005	0.008	0.005	0.005	0.026	0.004	0.007	
4	Politics-Raila	*Test Set*	0.00025	0.00025	0.00025	0.0003	0.00025	0.0003	0.0005	0.00079	0.00085	0.0009	**0.239**
		Control Set	0.0019	0.008	0.005	0.006	0.005	0.003	0.008	0.012	0.004	0.004	
5	Politics-Ruto	*Test Set*	0.0002	0.00025	0.00025	0.0002	0.00015	0.00018	0.0003	0.00055	0.00095	0.000115	**0.162**
		Control Set	0.005	0.016	0.005	0.007	0.010	0.011	0.0004	0.021	0.006	0.003	

Table 6.2: Sample sub-topics with corresponding semantic weights in test and control sets over a period of 10 quarters. Pearson Correlation Coefficient (PCC) between test and control sets for each subtopic per time stamp was computed depicting the semantic changes as validation.

Furthermore, the correlation in the test and control sets was evaluated by measuring the Pearson Correlation Coefficient (PCC) between the two sets at each timestamp (Benesty et al., 2009) across the 10 quarters. Ideally, there is a linearity assumption in the relationships between the weights. This is in the sense that an increase or decrease in the semantic topical weight of the test set should have the same effect as in the control set, depicting uniformity in interest over time. The coefficient is computed as the covariance of the two variables per timestamp divided by the product of the standard deviation in the control and test weight sets as $PCC = \mathrm{Cov}(a, b)/\sigma(a) * \sigma(b)$. Here, PCC is the Pearson's correlation coefficient and σ is the standard deviation that is applied to variables a and b. The correlation is expressed with a value between -1 and 1, where -1 depicts a negative correlation while 1 indicates positive correlation.

The results in Table 6.2 depict positive correlation between the semantic weights for both the control and test sets across the subtopics. It is worth noting that the most positive correlations were noted in emerging topics such as *"Diseases Related - Coronavirus"* and *"Diseases Related - Quarantine"*. The same positivity in the correlations is noted in the *political* and *current affairs* subtopics, albeit to a lesser extent. This is attributed to the sinusoidal pattern in the interest weights across the evaluation timestamps.

6.6 Discussion and Application Areas

The goal of the work in this chapter has been to identify evolving topical interests in short texts and eventually build user-representative profiles. Prior research in the evolution of interests in streaming platforms encompassed several factors. Usage of external data, annotation, and a mixture of features in the augmentation of user interests in the profiling process has been studied (Jiang & Sha, 2015; Zhu et al., 2019; Stai, Milaiou, Karyotis & Papavassiliou, 2018). Word embeddings representations have also been utilised in the generation of dynamic user profiles (Liang et al., 2018a). Inspired

by these insights, dynamicity in this approach was at topical level, where the inner product between word and topical embeddings representing each topical interest at different timestamps was derived.

The proposed representation model allows for the investigation of relationships between topics and word embeddings at each timestamp. To do so, each term is distributed over topical embeddings where each topic is denoted as a time-variational vector. Therefore, each word is probabilistically assigned to a topic. The word-topic correlation is higher in the semantic space, when the semantic relevance between the two is also high. With the consideration of time priors over the topical embeddings, variations of topical interests are captured over time.

To evaluate the modelling framework, both quantitative and qualitative measures were implemented. Quantitatively, topical qualities across the embeddings and baselines were computed. A product of topical diversity and coherence was the topical quality measure. FastText Skip-Gram variant gave the best results in terms of the topical quality thus was the word embedding algorithm of choice for further evaluations. Qualitatively, a measure of *intra-topical changes over time* and *topical interests' correlation* was evaluated. Topical terms with their respective probabilistic weights across the timestamps were generated with results depicted in Figures B.1, B.2, B.3, B.4, B.5, B.6, B.7, B.8, B.9 and B.10. Ideally, topical terms that generated attention had higher weights and this varied across the timestamps.

Furthermore, a control set was needed in depicting the correlation of interests with the test set. The idea was to correlate the semantic weights in the two sets. This way, it was possible to extract variations in the topical interests over the timestamps. The Pearson Correlation Coefficient (PCC) between the values in the two sets validated the semantic changes. The results are presented in Table 6.2 as well as in Figure 6.3. From the results, emerging topical words like "coronavirus" depicted greater correlations in the two sets. However, co-occurring terms across the timestamps, e.g., political interests,

fluctuated over time, a typical occurrence even in conventional news.

A few practical implications and application areas of the results are in the following spheres particularly relevant to third-party content providers as well as short-text data dissemination platforms: -

- Dissemination of semantically relevant time-variational content to users on the short-text platforms.
- Accuracy in the forecasts of the type/nature of content users are bound to consume in certain demographics and the likelihood of their future consumption by modelling their current consumption patterns.
- Content engagement patterns over time as content related to certain topics may be of more interest at specific times.
- Identification of the most relevant users to serve content by third-party content disseminators. This is relevant in cold-start scenarios too.

6.7 Chapter Summary

Microblogging platforms present intrinsic and extrinsic user profiles to third party content providers. This to a large extent is based on the nature of content that users and their friendship networks consume to a large or lesser extent over time. Thus, a time-dependent modelling approach is vital in accurate capturing of variational and user-representative profiles for short-text microblog users.

A modelling framework that factors variational timestamps on topical embeddings was proposed, implemented and tested in this chapter. FastText embeddings were used based on the algorithm's success in modelling short and noisy texts in Chapters 4 and 5. An ideal short-text microblogging scenario encompasses a factor of time. This

factors in the element of content gain and decay, a typical phenomenon in the platforms. Technically, quarterly timestamps were incorporated into the collection and modelling of the dataset. As much as this timestamp period was a bit long, it was bound to give a better representation of the topical evolvement process across the test and control sets. The test set was subdivided in terms of quarters from 2015 Quarter one (Q1) to 2020 Quarter two (Q2) and subjected through the framework. The resultant output was the semantic weight of subtopics within broader topics over 17 quarters in the test set.

To affirm the results in the test set, a control set was also setup. This set, as detailed in Section 6.4.1, was a corpus of news items collected at a time that reflected part of the test set collection time period, i.e., from 2018 Q1 to 2020 Q2. A correlation measure to determine the level of agreement in the semantic weights across the test and control sets was computed as the Pearson Correlation Coefficient (PCC) in each quarter. A PCC value close to 1 indicated better correlations between the topical relevance in the two sets at the specific timestamps.

In summary, the study in this chapter provided an enhanced modelling framework where time was a factor reminiscent of microblogging platforms. This sheds light on the design of third-party recommender systems in such short-text frameworks where interest gain and decay in user interests and successive profiling is vital. For instance, it was possible to deduce the interest changes in COVID-19 related content by tracking the dissemination patterns over time. The detailed approach has improved on the current state-of-the-art in time-variational topic modelling and interest's identification processes.

Chapter 7

Conclusion and Future Directions

7.1 Research Achievements

In this thesis, the problem of generating user-representative profiles in the design of third-party content recommendations in short-text microblogs was addressed. Key to addressing this personalisation problem, was the proposal and implementation of short-text modelling frameworks, as well as extraction of user-specific interests. Overall, the result was an exhaustive framework that consumed noisy and short texts, augmented them to an understandable format, at least by human interpretation standards, in addition to extracting semantic interests from the same. The end result was user-definitive profiles that incorporated evolvement patterns reminiscent of short-text microblog platforms. A summary of accomplishments attained in achieving each of these objectives is highlighted below.

1. **Chapter 3: Accurate extraction of notions/concepts using external data to augment short and noisy texts.**

   A novel short-text modelling framework called Metamodel LDA (MELDA) was presented which was an extension of LDA by incorporating metadata from an external but vocabulary-rich dataset. In the modelling process, a semantically

relevant long-text dataset was used to augment the text modelling process in the shorter texts. A two-step process was followed in the incorporation of longer texts in the modelling process. Firstly, topic labels were extracted from semantically relevant long texts and modelled via LDA, which works better with high vocabulary co-occurrences. Secondly, word tokens in the short texts were distributed over the topic labels extracted from the longer texts. This distribution happened at the LDA initialisation phase. The distribution process generated the (metamodel topic labels-tweets) matrix. The matrix computation process was guided by a pre-set seed confidence value as a percentage value. The seed value determined the restrain levels in the topical distribution patterns between the two datasets. If the terms of interest in the short-text topics fell within the set seed value over the long-text topics, the topical representation of the longer texts was assumed. Otherwise, the seed topics in the short texts took precedence. Overall, the MELDA framework extracted more interpretable notions/concepts from short texts.

2. **Chapter 4: User interest levels formulation in specific microblog content.** User-representative profiling largely relies on extraction of interests that define the users. Data modelling problems related to vocabulary sparsity, and lack of external data as detailed in Section 4.2 exist in short texts. Therefore, this prompted for the incorporation of neural network models for contextual knowledge extraction. Low vector representations were modelled to achieve this. To ascertain that the neural network model in use were semantically relevant, FastText (Skip Gram, CBOW), Word2Vec (Skip Gram, CBOW) and Glove were used to vectorise texts in the experimental framework.

   In extracting interests and computing user affinity towards the same, K-Means++

was used to extract cluster centroids representative of the interests. FastText Skip-Gram vectors generated the best set of clusters by Fowlkes-Mallows scores(FMI) and Silhouette Coefficient (S-Score) values. For each user (tweeter), a maximum of 3200 tweets were extracted from the users' timelines. The tweets were then pre-processed and vectorised. A cosine distance measure was applied to compute the distance of the vectorised tweets to the clusters of interest. The distance measure (Degree of Interest) depicting the affinity towards the specific clusters was assumed to be the interest level to the topical cluster. In the experimental framework, interests related to *Daily News Chatter*, *Swahili Content* and *Sports Betting* were computed. Average distances close to 0 indicated minimal interest while values close to 1 indicated higher affinity to the topics.

Quantitative and qualitative evaluation of the interest's formulation process was performed. A correlation measure among users and their friendship networks was also computed. A FastText Skip-Gram variant depicted positivity in the correlation measure. In addition, follow-back correlations for users in one common topical cluster were computed to ascertain the quantitative evaluation results. Overall, FastText based algorithms depicted interests and subsequent affinities better among all the compared models.

3. **Chapter 5: Diversity in user interest levels in short-text microblog content.** Diversity in content is characteristic of short-text microblog's dissemination and consumption patterns. A novel four step process framework was designed and experimented on a Twitter dataset. The goal was to profile users' tastes based on the diversification of their interests across several topical interests. Firstly, the short-text dataset was modelled via neural language models for contextual knowledge extraction. Secondly, the resultant vectors were subjected to a clustering process where the most contextually relevant short texts were grouped together

based on the most optimal interests from the dataset. Thirdly, a responsibility matrix was generated based on a set of test data identifying the interest levels to the topical interests using Expectation Maximisation (EM). Ideally, EM was definitive in the depiction of the extent to which the topical interests had over the test tweets. Fourthly, aggregation and averaging of individual EM weights in the responsibility matrix was then computed resulting in a user-interest representative model.

A Multi-interest user profile comprising of the averaged EM weights was the final output. The aggregated weights per user totalled to 1 as the individual weights were probabilistic. Thresholding of the interest weights in the results was a possibility. In this framework, the inter-topic interest median was applied as the threshold. Therefore, a user's topical interest was significant if it was equal to or greater than the threshold. To validate the results, a set of tweeters with known interests were subjected to the framework. The model's result qualitatively corroborated the findings of human validators in terms of the resultant Kappa scores between the model's and human's evaluation.

4. **Chapter 6: Design, implementation and evaluation of a time-variational user interest extraction and profiling framework.**

   Lastly, time variational modelling and profiling, reminiscent of short-text microblogs was explored. This was attributed to the volatility in the dissemination patterns and nature of content on such platforms. A factor of time was introduced in this modelling framework. In this case, the modelling process was segmented per specified periods. In each timestamp, topically representative terms over word embeddings were generated. The resultant output at each timestamp was the inner product of word and topic embeddings for better generalisation. In this inferencing phase, correlations between terms and respective topics of interest

were higher when the term and topic embedding were also high, like in clustering. Smoothness of topics over embeddings was enforced by Markov Chain.

The modelling framework was subjected to test and control/validation set of tweets for evaluation. Quantitatively, topical quality was measured across timestamps. FastText Skip-Gram variant generated quality topics by topical diversity and coherence compared to the other modelling approaches including baselines like Twitter-LDA. Qualitatively, intra-topical variations over time and topical interests' correlation in the test and control sets were measured. A Pearson Correlation Coefficient (PCC) between the sets over the test period was computed. A PCC value as high as 0.871 was achieved in emerging topics indicating the quality of the topical interests in comparison to the control set. The obtained results were significant in the design of short text-based recommender systems.

In summary, different approaches have been presented in the formulation of user-representative profiles based on short-text microblog data. An approach utilizing a topically relevant dataset is proposed to augment the knowledge extraction process from short and noisy texts. Follow-back recommendations based on specific user interests in short texts follows. To further refine the user interests modelling process, variations in the affinities towards user interests in the profiling process is measured. Lastly, time-based variational modelling is incorporated in the profiling process where interest as a factor of time is evaluated. These stepwise processes are representative of a multi-faceted profiling paradigm. Here, keywords and concepts/notions are key in the generation of short text-based user-representative profiles, a key component in the design of third-party content recommender systems.

7.2 Limitations and Future work

There are some limitations in the current study which can be promising directions for future research in this domain.

1. **Augmentation of short texts for versatility in especially multi-lingual modelling.**

   The bulk of the disseminated content on short-text microblogs is in demographically relevant languages. For example, use of contextual models like Bidirectional Encoder Representations from Transformers (BERT) to a large extent is likely to improve on contextual augmentation of short texts. As opposed to the usage of a semantically relevant long-text dataset, such larger and multi-lingual models can be used regardless of the semantic relevance. Auto-regressive language models like GPT-3 and related variants can also be utilised it this respect. With 175 billion parameters and its text auto-generating ability, extra related data and features can be generated from the model. Weights to different data aspects can be applied in augmenting the short texts for better modelling. Languages modelling is one area of interest in future. For example, tweets are largely demographically relevant. There is need to preserve their language context without relying on English data for better modelling. We conjecture that Many-to-Many multilingual translation models such as Facebook's M2M-100 (Fan et al., 2020; Schwenk, Wenzek, Edunov, Grave & Joulin, 2019; El-Kishky, Chaudhary, Guzman & Koehn, 2019) would be integral in the augmentation of short texts to fit certain profiling tasks in the future.

   In addition, techniques that consider both concepts of paraphrases and non-paraphrases as binary relations over short texts can be utilised in the augmentation of short texts for better modelling. This might involve extracting textual and related features in short texts via several approaches including usage of neural

networks. An evaluation of soft clustering in topical interests extraction based on these contextual models is a potential research direction in the future.

2. **Mapping extrinsic and intrinsic user interests**

   User interests are core in the generation of user-representative profiles. In this study, disseminated content was of essence. However, the number of lurkers on short-text microblogs is quite substantial. Therefore, better interest's extraction processes devoid of the disseminated content are needed for better profiling of such users who rarely disseminate data. A probable approach would be in integrating such user's extrinsic interests with their network's intrinsic ones. This way, it will be possible to leverage on the homophily theory in modelling their interests for eventual profiling. Hybrid content, and network leveraging techniques are ideal in such tasks.

3. **Integration of communities and sub-communities in the profiling frameworks.**

   Integration of communities on short-text microblogging platforms is bound to improve the overall modelling results. In this work, friendship connections were pivotal in several evaluation strategies. However, networks beyond high level friendship connections have the potential of advancing the profiling process. To put this in context, user communities and sub-communities are bound to evolve and assume other unique parameters with contextual variations over time. Integration of community detection algorithms in the modelling process with emphasis on time and content sensitivity is bound to generate more user-representative profiles that not only factor static data but evolving data dissemination patterns and friendship networks.

4. **Dynamicity and time as modelling factors in short-text microblogs.**

   Content and network dynamicity and time are modelling factors in extracting

evolving user interests for modelling in short-text microblogs. Dynamism in non-lurkers is evident in the disseminated content as well as in their networks. In the study, only content dynamism was dealt with in the profiling perspective. As part of the future work, network dynamism can be factored in the profiling process. In short-text follower-followee microblogging platforms, new users may join the network as current users leave the same network. The same is replicated when some users unfollow others. Therefore, intra-user/community relationships change rapidly thus are pertinent in profiling users. Exploring a combination of approaches related to heterogeneity and dynamic concept graphs is another promising direction in the future. The concept graph over time could be incorporated in the user interest's extraction process or in short-text augmentation.

References

Abbasi, M. A., Tang, J. & Liu, H. (2014). Scalable learning of users' preferences using networked data. In *Proceedings of the 25th acm conference on hypertext and social media* (pp. 4–12).

Abdel-Hafez, A. & Xu, Y. (2013). A survey of user modelling in social media websites. *Computer and Information Science*, *6*(4), 59–71.

Abel, F., Gao, Q., Houben, G.-J. & Tao, K. (2011a). Analyzing temporal dynamics in twitter profiles for personalized recommendations in the social web. In *Proceedings of the 3rd international web science conference* (pp. 1–8).

Abel, F., Gao, Q., Houben, G.-J. & Tao, K. (2011b). Analyzing user modeling on twitter for personalized news recommendations. In *international conference on user modeling, adaptation, and personalization* (pp. 1–12).

Abel, F., Gao, Q., Houben, G.-J. & Tao, K. (2011c). Analyzing user modeling on twitter for personalized news recommendations. In *international conference on user modeling, adaptation, and personalization* (pp. 1–12).

Abel, F., Gao, Q., Houben, G.-J. & Tao, K. (2011d). Semantic enrichment of twitter posts for user profile construction on the social web. In *Extended semantic web conference* (pp. 375–389).

Abel, F., Hauff, C., Houben, G.-J. & Tao, K. (2012). Leveraging user modeling on the social web with linked data. In *International conference on web engineering* (pp. 378–385).

Aggarwal, C. C. (2016). Content-based recommender systems. In *Recommender systems* (pp. 139–166). Springer.

Aggarwal, C. C. & Zhai, C. (2012). *Mining text data*. Springer Science & Business Media.

Ahmed, A., Low, Y., Aly, M., Josifovski, V. & Smola, A. J. (2011). Scalable distributed inference of dynamic user interests for behavioral targeting. In *Proceedings of the 17th acm sigkdd international conference on knowledge discovery and data mining* (pp. 114–122).

Akbari, M., Hu, X., Liqiang, N. & Chua, T.-S. (2016). From tweets to wellness: Wellness event detection from twitter streams. In *Thirtieth aaai conference on artificial intelligence.*

Alaoui, S., Idrissi, Y. E. B. E. & Ajhoun, R. (2015). Building rich user profile based on intentional perspective. *Procedia Computer Science*, *73*, 342–349.

Alsaeedi, A. (2020). A survey of term weighting schemes for text classification.

International Journal of Data Mining, Modelling and Management, *12*(2), 237–254.

Andrzejewski, D., Zhu, X. & Craven, M. (2009). Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *Proceedings of the 26th annual international conference on machine learning* (pp. 25–32). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1553374.1553378 doi: 10.1145/1553374.1553378

Arthur, D. & Vassilvitskii, S. (2006). *k-means++: The advantages of careful seeding* (Tech. Rep.).

Arthur, D. & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual acm-siam symposium on discrete algorithms* (pp. 1027–1035).

Banerjee, N., Chakraborty, D., Dasgupta, K., Mittal, S., Joshi, A., Nagar, S., ... Madan, S. (2009). User interests in social media sites: an exploration with micro-blogs. In *Proceedings of the 18th acm conference on information and knowledge management* (pp. 1823–1826).

Bao, H., Li, Q., Liao, S. S., Song, S. & Gao, H. (2013). A new temporal and social pmf-based method to predict users' interests in micro-blogging. *Decision Support Systems*, *55*(3), 698–709.

Benesty, J., Chen, J., Huang, Y. & Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing* (pp. 1–4). Springer.

Ben-Lhachemi, N. et al. (2018). Using tweets embeddings for hashtag recommendation in twitter. *Procedia Computer Science*, *127*, 7–15.

Besag, J. (2004). An introduction to markov chain monte carlo methods. In *Mathematical foundations of speech and language processing* (pp. 247–270). Springer.

Besel, C., Schlötterer, J. & Granitzer, M. (n.d.). Inferring semantic interest profiles from twitter followees.

Bhargava, P., Brdiczka, O. & Roberts, M. (2015). Unsupervised modeling of users' interests from their facebook profiles and activities. In *Proceedings of the 20th international conference on intelligent user interfaces* (pp. 191–201).

Bhattacharya, P., Zafar, M. B., Ganguly, N., Ghosh, S. & Gummadi, K. P. (2014). Inferring user interests in the twitter social network. In *Proceedings of the 8th acm conference on recommender systems* (pp. 357–360).

Bholowalia, P. & Kumar, A. (2014). Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, *105*(9).

Blei, D. M. & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on machine learning* (pp. 113–120).

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003a). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003b, March). Latent dirichlet allocation. *J. Mach. Learn. Res.*, *3*, 993–1022. Retrieved from http://dl.acm.org/citation.cfm?id=944919.944937

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2016). Enriching word vectors

with subword information. *CoRR*, *abs/1607.04606*. Retrieved from `http://arxiv.org/abs/1607.04606`

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T. & Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 acm sigmod international conference on management of data* (pp. 1247–1250).

Bontcheva, K. & Rout, D. (2014). Making sense of social media streams through semantics: a survey. *Semantic Web*, *5*(5), 373–403.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 31–40.

Boyd, D. M. & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of computer-mediated communication*, *13*(1), 210–230.

Brena, G., Brambilla, M., Ceri, S., Di Giovanni, M., Pierri, F. & Ramponi, G. (2019). News sharing user behaviour on twitter: A comprehensive data collection of news articles and social interactions. In *Proceedings of the international aaai conference on web and social media* (Vol. 13, pp. 592–597).

Brusilovsky, P. & Millán, E. (2007). User models for adaptive hypermedia and adaptive educational systems. In *The adaptive web* (pp. 3–53). Springer.

Budak, C., Kannan, A., Agrawal, R. & Pedersen, J. (2014). Inferring user interests from microblogs. *AAAI ICWSM*.

Cai, Y., Leung, H.-f., Li, Q., Min, H., Tang, J. & Li, J. (2013). Typicality-based collaborative filtering recommendation. *IEEE Transactions on Knowledge and Data Engineering*, *26*(3), 766–779.

Cami, B. R., Hassanpour, H. & Mashayekhi, H. (2019). User preferences modeling using dirichlet process mixture model for a content-based recommender system. *Knowledge-Based Systems*, *163*, 644–655.

Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.-Y. & Moon, S. (2007). I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Proceedings of the 7th acm sigcomm conference on internet measurement* (pp. 1–14).

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L. & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems* (pp. 288–296).

Chapelle, O., Scholkopf, B. & Zien, A. (2009). Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, *20*(3), 542–542.

Chen, C. & Ren, J. (2017). Forum latent dirichlet allocation for user interest discovery. *Knowledge-Based Systems*, *126*, 1–7.

Chen, J., Nairn, R., Nelson, L., Bernstein, M. & Chi, E. (2010). Short and tweet: experiments on recommending content from information streams. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1185–1194).

Chen, K., Chen, T., Zheng, G., Jin, O., Yao, E. & Yu, Y. (2012). Collaborative personalized tweet recommendation. In *Proceedings of the 35th international acm sigir conference on research and development in information retrieval* (pp. 661–670).

Chen, M., Ghorbani, A. A. et al. (2019). A survey on user profiling model for anomaly detection in cyberspace. *Journal of Cyber Security and Mobility*, *8*(1), 75–112.

Chen, Z., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M. & Ghosh, R. (2013a). Discovering coherent topics using general knowledge. In *Proceedings of the 22nd acm international conference on information & knowledge management* (pp. 209–218). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2505515.2505519 doi: 10.1145/2505515.2505519

Chen, Z., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M. & Ghosh, R. (2013b). Leveraging multi-domain prior knowledge in topic models. In *Ijcai* (Vol. 13, pp. 2071–2077).

Chen, Z., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M. & Ghosh, R. (2013c). Leveraging multi-domain prior knowledge in topic models. In *Proceedings of the twenty-third international joint conference on artificial intelligence* (pp. 2071–2077). AAAI Press. Retrieved from http://dl.acm.org/citation.cfm?id=2540128.2540426

Chi, E. H. (2008). The social web: Research and opportunities. *Computer*(9), 88–91.

Chowdhury, J. R., Caragea, C. & Caragea, D. (2020). On identifying hashtags in disaster twitter data. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 34, pp. 498–506).

Collins, A. M. & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological review*, *82*(6), 407.

Cui, R., Agrawal, G. & Ramnath, R. (2020). Tweets can tell: activity recognition using hybrid gated recurrent neural networks. *Social Network Analysis and Mining*, *10*(1), 1–15.

Cunningham, P., Cord, M. & Delany, S. J. (2008). Supervised learning. In M. Cord & P. Cunningham (Eds.), *Machine learning techniques for multimedia: Case studies on organization and retrieval* (pp. 21–49). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from https://doi.org/10.1007/978-3-540-75171-7_2 doi: 10.1007/978-3-540-75171-7_2

Dantzig, G. B. & Veinott, A. F. (1968). *Mathematics of the decision sciences* (Vol. 2). American Mathematical Soc.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Dhelim, S., Aung, N. & Ning, H. (2020). Mining user interest based on personality-aware hybrid filtering in social networks. *Knowledge-Based Systems*, *206*, 106227.

Dieng, A. B., Ruiz, F. J. & Blei, D. M. (2019a). The dynamic embedded topic model. *arXiv preprint arXiv:1907.05545*.

Dieng, A. B., Ruiz, F. J. & Blei, D. M. (2019b). Topic modeling in embedding spaces. *arXiv preprint arXiv:1907.04907*.

Dieng, A. B., Ruiz, F. J. & Blei, D. M. (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, *8*, 439–453.

Di Tommaso, G., Faralli, S., Stilo, G. & Velardi, P. (2018). Wiki-mid: a very large multi-domain interests dataset of twitter users with mappings to wikipedia. In *International semantic web conference* (pp. 36–52).

Dong, A., Zhang, R., Kolari, P., Bai, J., Diaz, F., Chang, Y., ... Zha, H. (2010). Time is of the essence: Improving recency ranking using twitter data. In *Proceedings of the 19th international conference on world wide web* (p. 331–340). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/1772690.1772725 doi: 10.1145/1772690.1772725

Dooley, P. & Božić, B. (2019). Towards linked data for wikidata revisions and twitter trending hashtags. In *Proceedings of the 21st international conference on information integration and web-based applications & services* (pp. 166–175).

El-Arini, K., Paquet, U., Herbrich, R., Van Gael, J. & Agüera y Arcas, B. (2012). Transparent user models for personalization. In *Proceedings of the 18th acm sigkdd international conference on knowledge discovery and data mining* (pp. 678–686).

El-Kishky, A., Chaudhary, V., Guzman, F. & Koehn, P. (2019). A massive collection of cross-lingual web-document pairs. *arXiv preprint arXiv:1911.06154*.

Elmongui, H. G., Mansour, R., Morsy, H., Khater, S., El-Sharkasy, A. & Ibrahim, R. (2015). Trupi: Twitter recommendation based on users' personal interests. In *International conference on intelligent text processing and computational linguistics* (pp. 272–284).

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., ... Joulin, A. (2020). Beyond english-centric multilingual machine translation. *arXiv preprint*.

Faralli, S., Stilo, G. & Velardi, P. (2015). Recommendation of microblog users based on hierarchical interest profiles. *Social Network Analysis and Mining*, *5*(1), 25.

Faralli, S., Stilo, G. & Velardi, P. (2017). Automatic acquisition of a taxonomy of microblogs users' interests. *Journal of Web Semantics*, *45*, 23–40.

Färber, M., Ell, B., Menne, C. & Rettinger, A. (2015). A comparative survey of dbpedia, freebase, opencyc, wikidata, and yago. *Semantic Web Journal*, *1*(1), 1–5.

Farseev, A., Akbari, M., Samborskii, I. & Chua, T.-S. (2016). " 360° user profiling: past, future, and applications" by aleksandr farseev, mohammad akbari, ivan samborskii and tat-seng chua with martin vesely as coordinator. *ACM SIGWEB Newsletter*(Summer), 1–11.

Farseev, A., Nie, L., Akbari, M. & Chua, T.-S. (2015). Harvesting multiple sources for user profile learning: A big data study. In *Proceedings of the 5th acm on international conference on multimedia retrieval* (p. 235–242). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/2671188.2749381 doi: 10.1145/2671188.2749381

Figueiredo, F. & Jorge, A. (2019). Identifying topic relevant hashtags in twitter streams. *Information Sciences*, *505*, 65–83.

Flati, T., Vannella, D., Pasini, T. & Navigli, R. (2014). Two is bigger (and better) than one: the wikipedia bitaxonomy project. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 945–955).

Gabrilovich, E., Markovitch, S. et al. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Ijcai* (Vol. 7, pp. 1606–1611).

Gao, Q., Abel, F., Houben, G.-J. & Tao, K. (2011). Interweaving trend and user modeling for personalized news recommendation. In *2011 ieee/wic/acm international conferences on web intelligence and intelligent agent technology* (Vol. 1, pp. 100–103).

Garcia Esparza, S., O'Mahony, M. P. & Smyth, B. (2013). Catstream: categorising tweets for user profiling and stream filtering. In *Proceedings of the 2013 international conference on intelligent user interfaces* (pp. 25–36).

Gauch, S., Speretta, M., Chandramouli, A. & Micarelli, A. (2007). User profiles for personalized information access. In *The adaptive web* (pp. 54–89). Springer.

Goel, S. & Kumar, R. (2018). Folksonomy-based user profile enrichment using clustering and community recommended tags in multiple levels. *Neurocomputing*, *315*, 425–438.

Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, *10*(1), 1–309.

Grbovic, M., Radosavljevic, V., Djuric, N., Bhamidipati, N. & Nagarajan, A. (2016). Leveraging blogging activity on tumblr to infer demographics and interests of users for advertising purposes. In *# microposts* (pp. 2–11).

Grenha Teixeira, J., Patrício, L., Huang, K.-H., Fisk, R. P., Nóbrega, L. & Constantine, L. (2017). The minds method: integrating management and interaction design perspectives for service design. *Journal of Service Research*, *20*(3), 240–258.

Griffiths, T. L. & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*(suppl 1), 5228–5235.

Große-Bölting, G., Nishioka, C. & Scherp, A. (2015). Generic process for extracting user profiles from social media using hierarchical knowledge bases. In *Proceedings of the 2015 ieee 9th international conference on semantic computing (ieee icsc 2015)* (pp. 197–200).

Halberstam, Y. & Knight, B. (2016). Homophily, group size, and the diffusion of political information in social networks: Evidence from twitter. *Journal of Public Economics*, *143*, 73–88.

Hamdi, T., Slimi, H., Bounhas, I. & Slimani, Y. (2020). A hybrid approach for fake news detection in twitter based on user features and graph embedding. In *International conference on distributed computing and internet technology* (pp. 266–280).

Han, J. & Pei, J. (2000). Mining frequent patterns by pattern-growth: methodology and implications. *ACM SIGKDD explorations newsletter*, *2*(2), 14–20.

Hannak, A., Margolin, D., Keegan, B. & Weber, I. (2014). Get back! you don't know me like that: The social mediation of fact checking interventions in twitter conversations. In *Eighth international aaai conference on weblogs and social media.*

Hannon, J., McCarthy, K., O'Mahony, M. P. & Smyth, B. (2012). A multi-faceted user model for twitter. In *International conference on user modeling, adaptation, and personalization* (pp. 303–309).

Hassan, N., Arslan, F., Li, C. & Tremayne, M. (2017). Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1803–1812).

He, Y., Wang, C. & Jiang, C. (2015). Discovering canonical correlations between topical and topological information in document networks. In *Proceedings of the 24th acm international on conference on information and knowledge management* (pp. 1281–1290). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2806416.2806518 doi: 10.1145/2806416.2806518

He, Y., Wang, C. & Jiang, C. (2017a). Mining coherent topics with pre-learned interest knowledge in twitter. *IEEE Access*, *5*, 10515–10525.

He, Y., Wang, C. & Jiang, C. (2017b). Modeling document networks with tree-averaged copula regularization. In *Proceedings of the tenth acm international conference on web search and data mining* (pp. 691–699). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/3018661.3018666 doi: 10.1145/3018661.3018666

He, Y., Wang, C. & Jiang, C. (2017c). Multi-perspective hierarchical dirichlet process for geographical topic modeling. In *Pacific-asia conference on knowledge discovery and data mining* (pp. 811–823).

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international acm sigir conference on research and development in information retrieval* (pp. 50–57). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/312624.312649 doi: 10.1145/312624.312649

Iwata, T., Watanabe, S., Yamada, T. & Ueda, N. (2009). Topic tracking model for analyzing consumer purchase behavior. In *Twenty-first international joint conference on artificial intelligence.*

Jagarlamudi, J., Daumé III, H. & Udupa, R. (2012). Incorporating lexical priors into topic models. In *Proceedings of the 13th conference of the european chapter of the association for computational linguistics* (pp. 204–213).

Jawaheer, G., Weller, P. & Kostkova, P. (2014). Modeling user preferences in recommender systems: A classification framework for explicit and implicit user feedback. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *4*(2), 8.

Jiang, B. & Sha, Y. (2015). Modeling temporal dynamics of user interests in online social networks. *Procedia Computer Science*, *51*, 503–512.

Joshi, D., Cooper, M., Chen, F. & Chen, Y.-y. (2015). Building user profiles from shared photos. In *Proceedings of the 2015 workshop on community-organized multimodal mining: Opportunities for novel solutions* (pp. 37–42).

Kang, J. & Lee, H. (2017). Modeling user interest in social media using news media and wikipedia. *Information Systems*, *65*, 52–64.

Kanoje, S., Girase, S. & Mukhopadhyay, D. (2015). User profiling trends, techniques and applications. *arXiv preprint arXiv:1503.07474*.

Kapanipathi, P., Jain, P., Venkataramani, C. & Sheth, A. (2014a). Hierarchical interest graph from tweets. In *Proceedings of the 23rd international conference on world wide web* (pp. 311–312).

Kapanipathi, P., Jain, P., Venkataramani, C. & Sheth, A. (2014b). User interests identification on twitter using a hierarchical knowledge base. In *European semantic web conference* (pp. 99–113).

Kapanipathi, P., Orlandi, F., Sheth, A. P. & Passant, A. (2011). Personalized filtering of the twitter stream.

Karatay, D. & Karagoz, P. (2015). User interest modeling in twitter with named entity recognition. In *5th workshop on making sense of microposts.*

Karidi, D. P., Stavrakas, Y. & Vassiliou, Y. (2018). Tweet and followee personalized recommendations based on knowledge graphs. *Journal of Ambient Intelligence and Humanized Computing*, *9*(6), 2035–2049.

Khater, S., Elmongui, H. G. & Gracanin, D. (2014). Tweets you like: Personalized tweets recommendation based on dynamic users interests. In *Proc. int. conf. social comput.(socialcom)* (pp. 14–15).

Korenčić, D., Ristov, S. & Šnajder, J. (2018). Document-based topic coherence measures for news media text. *Expert Systems with Applications*, *114*, 357–373.

Korula, N. & Lattanzi, S. (2014). An efficient reconciliation algorithm for social networks. *Proceedings of the VLDB Endowment*, *7*(5), 377–388.

Kraut, R. E. & Resnick, P. (2012). *Building successful online communities: Evidence-based social design*. Mit Press.

Kwak, H., Lee, C., Park, H. & Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on world wide web* (pp. 591–600).

Landauer, T. K., Foltz, P. W. & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, *25*(2-3), 259–284.

Lapão, L. V., Da Silva, M. M. & Gregório, J. (2017). Implementing an online pharmaceutical service using design science research. *BMC medical informatics and decision making*, *17*(1), 31.

Li, J., Xu, H., He, X., Deng, J. & Sun, X. (2016). Tweet modeling with lstm recurrent neural networks for hashtag recommendation. In *2016 international joint conference on neural networks (ijcnn)* (pp. 1570–1577).

Li, Y. & Yuan, Y. (2017). Convergence analysis of two-layer neural networks with relu activation. In *Advances in neural information processing systems* (pp. 597–607).

Liang, S., Zhang, X., Ren, Z. & Kanoulas, E. (2018a). Dynamic embeddings for user profiling in twitter. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining* (pp. 1764–1773).

Liang, S., Zhang, X., Ren, Z. & Kanoulas, E. (2018b). Dynamic embeddings for user profiling in twitter. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining* (pp. 1764–1773). New

York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/3219819.3220043 doi: 10.1145/3219819.3220043

Liao, Y., Moshtaghi, M., Han, B., Karunasekera, S., Kotagiri, R., Baldwin, T., ... Pattison, P. (2012). Mining micro-blogs: Opportunities and challenges. In *Computational social networks* (pp. 129–159). Springer.

Liu, H., Hu, Z., Mian, A., Tian, H. & Zhu, X. (2014). A new user similarity model to improve the accuracy of collaborative filtering. *Knowledge-Based Systems*, *56*, 156–166.

Liu, P., Zhang, L. & Gulla, J. A. (2019). Real-time social recommendation based on graph embedding and temporal context. *International Journal of Human-Computer Studies*, *121*, 58–72.

Liu, Y., Chen, X., Li, S. & Wang, L. (2016). A user adaptive model for followee recommendation on twitter. In *Natural language understanding and intelligent applications* (pp. 425–436). Springer.

Liu, Z., Huang, W., Zheng, Y. & Sun, M. (2010). Automatic keyphrase extraction via topic decomposition. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 366–376).

Lu, C., Lam, W. & Zhang, Y. (2012). Twitter user modeling and tweets recommendation based on wikipedia concept graph. In *Workshops at the twenty-sixth aaai conference on artificial intelligence.*

Maimon, O. & Rokach, L. (2005). Data mining and knowledge discovery handbook.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, *22*(3), 276–282.

McPherson, M., Smith-Lovin, L. & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, *27*(1), 415–444.

Mei, Q., Cai, D., Zhang, D. & Zhai, C. (2008). Topic modeling with network regularization. In *Proceedings of the 17th international conference on world wide web* (pp. 101–110). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1367497.1367512 doi: 10.1145/1367497.1367512

Michelson, M. & Macskassy, S. A. (2010). Discovering users' topics of interest on twitter: a first look. In *Proceedings of the fourth workshop on analytics for noisy unstructured text data* (pp. 73–80).

Mihalcea, R. & Tarau, P. (2004). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing* (pp. 404–411).

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, *38*(11), 39–41.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M. & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 262–272). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from `http://dl.acm.org/citation.cfm?id=2145432.2145462`

Mishra, S., Rizoiu, M.-A. & Xie, L. (2018). Modeling popularity in asynchronous social media streams with recurrent neural networks. In *Twelfth international aaai conference on web and social media.*

Mislove, A., Viswanath, B., Gummadi, K. P. & Druschel, P. (2010). You are who you know: inferring user profiles in online social networks. In *Proceedings of the third acm international conference on web search and data mining* (pp. 251–260).

Moody, C. E. (2016). Mixing dirichlet topic models and word embeddings to make lda2vec. *arXiv preprint arXiv:1605.02019*.

Narducci, F., Musto, C., Semeraro, G., Lops, P. & De Gemmis, M. (2013). Leveraging encyclopedic knowledge for transparent and serendipitous user profiles. In *International conference on user modeling, adaptation, and personalization* (pp. 350–352).

Nguyen, V.-D., Sriboonchitta, S. & Huynh, V.-N. (2017). Using community preference for overcoming sparsity and cold-start problems in collaborative filtering system offering soft ratings. *Electronic Commerce Research and Applications*, *26*, 101–108.

Nishioka, C. & Scherp, A. (2016). Profiling vs. time vs. content: What does matter for top-k publication recommendation based on twitter profiles? In *2016 ieee/acm joint conference on digital libraries (jcdl)* (pp. 171–180).

Orlandi, F., Breslin, J. & Passant, A. (2012). Aggregated, interoperable and multi-domain user profiles for the social web. In *Proceedings of the 8th international conference on semantic systems* (pp. 41–48).

Ottoni, R., Las Casas, D., Pesce, J. P., Meira Jr, W., Wilson, C., Mislove, A. & Almeida, V. (2014). Of pins and tweets: Investigating how users behave across image-and text-based social networks. In *Eighth international aaai conference on weblogs and social media.*

Ouaftouh, S., Zellou, A. & Idri, A. (2015). User profile model: A user dimension based classification. In *2015 10th international conference on intelligent systems: Theories and applications (sita)* (pp. 1–5).

Palsetia, D., Patwary, M. M. A., Agrawal, A. & Choudhary, A. (2014). Excavating social circles via user interests. *Social Network Analysis and Mining*, *4*(1), 170.

Pathak, N., DeLong, C., Banerjee, A. & Erickson, K. (2008). Social topic models for community extraction. In *The 2nd sna-kdd workshop* (Vol. 8).

Paul, I., Khattar, A., Kumaraguru, P., Gupta, M. & Chopra, S. (2019). Elites tweet? characterizing the twitter verified user network. In *2019 ieee 35th international conference on data engineering workshops (icdew)* (pp. 278–285).

Peñas, P., Del Hoyo, R., Vea-Murguía, J., González, C. & Mayo, S. (2013). Collective knowledge ontology user profiling for twitter–automatic user profiling. In *2013 ieee/wic/acm international joint conferences on web intelligence (wi) and*

intelligent agent technologies (iat) (Vol. 1, pp. 439–444).

Pennacchiotti, M., Silvestri, F., Vahabi, H. & Venturini, R. (2012). Making your interests follow you on twitter. In *Proceedings of the 21st acm international conference on information and knowledge management* (pp. 165–174).

Pennington, J., Socher, R. & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543).

Pereira, F. S., Gama, J., de Amo, S. & Oliveira, G. M. (2018). On analyzing user preference dynamics with temporal social networks. *Machine Learning*, *107*(11), 1745–1773.

Peters, G., Crespo, F., Lingras, P. & Weber, R. (2013). Soft clustering–fuzzy and rough approaches and their extensions and derivatives. *International Journal of Approximate Reasoning*, *54*(2), 307–322.

Piao, G. & Breslin, J. G. (2016a). Exploring dynamics and semantics of user interests for user modeling on twitter for link recommendations. In *proceedings of the 12th international conference on semantic systems* (pp. 81–88).

Piao, G. & Breslin, J. G. (2016b). Interest representation, enrichment, dynamics, and propagation: a study of the synergetic effect of different user modeling dimensions for personalized recommendations on twitter. In *European knowledge acquisition workshop* (pp. 496–510).

Piao, G. & Breslin, J. G. (2017a). Inferring user interests for passive users on twitter by leveraging followee biographies. In *European conference on information retrieval* (pp. 122–133).

Piao, G. & Breslin, J. G. (2017b). Leveraging followee list memberships for inferring user interests for passive users on twitter. In *Proceedings of the 28th acm conference on hypertext and social media* (pp. 155–164).

Piao, G. & Breslin, J. G. (2018). Inferring user interests in microblogging social networks: a survey. *User Modeling and User-Adapted Interaction*, *28*(3), 277–329.

Piña-García, C., Siqueiros-García, J. M., Robles-Belmont, E., Carreón, G., Gershenson, C. & López, J. A. D. (n.d.). From neuroscience to computer science: a topical approach on twitter. *Journal of Computational Social Science*, 1–22.

Poo, D., Chng, B. & Goh, J.-M. (2003). A hybrid approach for user profiling. In *36th annual hawaii international conference on system sciences, 2003. proceedings of the* (pp. 9–pp).

Ramage, D., Dumais, S. & Liebling, D. (2010). Characterizing microblogs with topic models. In *Fourth international aaai conference on weblogs and social media.*

Ramage, D., Hall, D., Nallapati, R. & Manning, C. D. (2009). Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 conference on empirical methods in natural language processing: Volume 1 - volume 1* (pp. 248–256). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from `http://dl.acm.org/citation.cfm?id=1699510.1699543`

Ramasamy, D., Venkateswaran, S. & Madhow, U. (2013). Inferring user interests from

tweet times. In *Proceedings of the first acm conference on online social networks* (pp. 235–240).

Recalde, L. & Baeza-Yates, R. (2018). What kind of content are you prone to tweet? multi-topic preference model for tweeters. *arXiv preprint arXiv:1807.07162*.

Recalde, L. & Kaskina, A. (2017). Who is suitable to be followed back when you are a twitter interested in politics? In *Proceedings of the 18th annual international conference on digital government research* (pp. 94–99).

Röder, M., Both, A. & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth acm international conference on web search and data mining* (pp. 399–408).

Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. d. F. & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PloS one*, *14*(1), e0210236.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, 53–65.

Sahoo, S. R. & Gupta, B. (2019). Hybrid approach for detection of malicious profiles in twitter. *Computers & Electrical Engineering*, *76*, 65–81.

Sakaki, T., Okazaki, M. & Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on world wide web* (pp. 851–860).

Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*(5), 513–523.

Sang, J., Lu, D. & Xu, C. (2015). A probabilistic framework for temporal user modeling on microblogs. In *Proceedings of the 24th acm international on conference on information and knowledge management* (pp. 961–970).

Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, M. D. & Sperling, J. (2009). Twitterstand: news in tweets. In *Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems* (pp. 42–51).

Sasaki, K., Yoshikawa, T. & Furuhashi, T. (2014). Online topic model for twitter considering dynamics of user interests and topic trends. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1977–1985).

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., ... Lin, C.-T. (2017). A review of clustering techniques and developments. *Neurocomputing*, *267*, 664–681.

Schwenk, H., Wenzek, G., Edunov, S., Grave, E. & Joulin, A. (2019). Ccmatrix: Mining billions of high-quality parallel sentences on the web. *arXiv preprint arXiv:1911.04944*.

Shin, Y., Ryo, C. & Park, J. (2014). Automatic extraction of persistent topics from social text streams. *World Wide Web*, *17*(6), 1395–1420.

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H. & Demirbas, M. (2010). Short text classification in twitter to improve information filtering. In *Proceedings*

of the 33rd international acm sigir conference on research and development in information retrieval (pp. 841–842).

Stai, E., Milaiou, E., Karyotis, V. & Papavassiliou, S. (2018). Temporal dynamics of information diffusion in twitter: Modeling and experimentation. *IEEE Transactions on Computational Social Systems*, *5*(1), 256-264. doi: 10.1109/TCSS.2017.2784184

Steyvers, M. & Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis*, *427*(7), 424–440.

Suchanek, F. M., Kasneci, G. & Weikum, G. (2007). Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on world wide web* (pp. 697–706).

Symeonidis, P., Nanopoulos, A. & Manolopoulos, Y. (2007). Feature-weighted user model for recommender systems. In *International conference on user modeling* (pp. 97–106).

Tajbakhsh, M. S. & Bagherzadeh, J. (2019). Semantic knowledge lda with topic vector for recommending hashtags: Twitter use case. *Intelligent Data Analysis*, *23*(3), 609–622.

Takimura, S., Harakawa, R., Ogawa, T. & Haseyama, M. (2018). Twitter followee recommendation based on multimodal ffm considering social relations. In *2018 ieee 7th global conference on consumer electronics (gcce)* (pp. 204–205).

Tang, J., Hu, X. & Liu, H. (2013). Social recommendation: a review. *Social Network Analysis and Mining*, *3*(4), 1113–1133.

Trikha, A. K., Zarrinkalam, F. & Bagheri, E. (2018). Topic-association mining for user interest detection. In *European conference on information retrieval* (pp. 665–671).

van den Beukel, S., Goos, S. H. & Treur, J. (2019). An adaptive temporal-causal network model for social networks based on the homophily and more-becomes-more principle. *Neurocomputing*, *338*, 361–371.

Viera, A. J., Garrett, J. M. et al. (2005). Understanding interobserver agreement: the kappa statistic. *Fam med*, *37*(5), 360–363.

Vu, T. & Perez, V. (2013). Interest mining from user tweets. In *Proceedings of the 22nd acm international conference on information & knowledge management* (pp. 1869–1872).

Wandabwa, H., Naeem, M. A., Mirza, F. & Pears, R. (2020). Follow-back recommendations for sports bettors: A twitter-based approach. In *Proceedings of the 53rd hawaii international conference on system sciences.*

Wandabwa, H., Naeem, M. A., Mirza, F., Pears, R. & Nguyen, A. (2020). Multi-interest user profiling in short text microblogs. In S. Hofmann, O. Müller & M. Rossi (Eds.), *Designing for digital transformation. co-creating services with citizens and industry* (pp. 154–168). Cham: Springer International Publishing.

Wang, J., Zhao, W. X., He, Y. & Li, X. (2014). Infer user interests via link structure regularization. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *5*(2), 1–22.

Wang, T., Viswanath, V. & Chen, P. (2015). Extended topic model for word dependency. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 2: Short papers)* (Vol. 2, pp. 506–510).

Wang, X., Wei, F., Liu, X., Zhou, M. & Zhang, M. (2011). Topic sentiment analysis in twitter: A graph-based hashtag sentiment classification approach. In *Proceedings of the 20th acm international conference on information and knowledge management* (pp. 1031–1040). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2063576.2063726 doi: 10.1145/2063576.2063726

Wei, Y., Cheng, Z., Yu, X., Zhao, Z., Zhu, L. & Nie, L. (2019). Personalized hashtag recommendation for micro-videos. In *Proceedings of the 27th acm international conference on multimedia* (pp. 1446–1454).

Weng, J., Lim, E.-P., Jiang, J. & He, Q. (2010a). Twitterrank: Finding topic-sensitive influential twitterers. In *Proceedings of the third acm international conference on web search and data mining* (pp. 261–270). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1718487.1718520 doi: 10.1145/1718487.1718520

Weng, J., Lim, E.-P., Jiang, J. & He, Q. (2010b). Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third acm international conference on web search and data mining* (pp. 261–270).

Witten, I. H. & Frank, E. (2002). Data mining: practical machine learning tools and techniques with java implementations. *Acm Sigmod Record*, *31*(1), 76–77.

Xu, S. & Zhou, A. (2020). Hashtag homophily in twitter network: Examining a controversial cause-related marketing campaign. *Computers in Human Behavior*, *102*, 87–96.

Yang, L., Sun, T., Zhang, M. & Mei, Q. (2012). We know what@ you# tag: does the dual role affect hashtag adoption? In *Proceedings of the 21st international conference on world wide web* (pp. 261–270).

Yang, S.-H., Kolcz, A., Schlaikjer, A. & Gupta, P. (2014a). Large-scale high-precision topic modeling on twitter. In *Proceedings of the 20th acm sigkdd international conference on knowledge discovery and data mining* (pp. 1907–1916).

Yang, S.-H., Kolcz, A., Schlaikjer, A. & Gupta, P. (2014b). Large-scale high-precision topic modeling on twitter. In *Proceedings of the 20th acm sigkdd international conference on knowledge discovery and data mining* (pp. 1907–1916). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2623330.2623336 doi: 10.1145/2623330.2623336

yeon Sung, Y. & Kim, S. B. (2020). Topical keyphrase extraction with hierarchical semantic networks. *Decision Support Systems*, *128*, 113163.

Yin, H., Cui, B., Chen, L., Hu, Z. & Zhou, X. (2015a). Dynamic user modeling in social media systems. *ACM Transactions on Information Systems (TOIS)*, *33*(3), 1–44.

Yin, H., Cui, B., Chen, L., Hu, Z. & Zhou, X. (2015b, March). Dynamic user modeling in social media systems. *ACM Trans. Inf. Syst.*, *33*(3). Retrieved from

https://doi.org/10.1145/2699670 doi: 10.1145/2699670

Ying, Q. F., Chiu, D. M., Venkatramanan, S. & Zhang, X. (2018). User modeling and usage profiling based on temporal posting behavior in osns. *Online Social Networks and Media*, *8*, 32–41.

Yu, D., Xu, D., Wang, D. & Ni, Z. (2019). Hierarchical topic modeling of twitter data for online analytical processing. *IEEE Access*, *7*, 12373–12385.

Zafarani, R. & Liu, H. (2013). Connecting users across social media sites: a behavioral-modeling approach. In *Proceedings of the 19th acm sigkdd international conference on knowledge discovery and data mining* (pp. 41–49).

Zarrinkalam, F., Fani, H. & Bagheri, E. (2019). Social user interest mining: Methods and applications. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining* (pp. 3235–3236).

Zarrinkalam, F., Fani, H., Bagheri, E. & Kahani, M. (2017). Predicting users' future interests on twitter. In *European conference on information retrieval* (pp. 464–476).

Zarrinkalam, F., Fani, H., Bagheri, E., Kahani, M. & Du, W. (2015). Semantics-enabled user interest detection from twitter. In *2015 ieee/wic/acm international conference on web intelligence and intelligent agent technology (wi-iat)* (Vol. 1, pp. 469–476).

Zhan, J. & Dahal, B. (2017). Using deep learning for short text understanding. *Journal of Big Data*, *4*(1), 34.

Zhang, Z., Robinson, D. & Tepper, J. (2018). Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference* (pp. 745–760).

Zhao, W. X., Jiang, J., He, J., Song, Y., Achananuparp, P., Lim, E.-P. & Li, X. (2011). Topical keyphrase extraction from twitter. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies - volume 1* (pp. 379–388). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=2002472.2002521

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H. & Li, X. (2011). Comparing twitter and traditional media using topic models. In *European conference on information retrieval* (pp. 338–349).

Zhao, Z., Cheng, Z., Hong, L. & Chi, E. H. (2015). Improving user topic interest profiles by behavior factorization. In *Proceedings of the 24th international conference on world wide web* (pp. 1406–1416).

Zheng, J., Wang, S., Li, D. & Zhang, B. (2019). Personalized recommendation based on hierarchical interest overlapping community. *Information Sciences*, *479*, 55–75.

Zheng, X. & Sun, A. (2019). Collecting event-related tweets from twitter stream. *Journal of the Association for Information Science and Technology*, *70*(2), 176–186.

Zhou, X., Xu, Y., Li, Y., Josang, A. & Cox, C. (2012). The state-of-the-art in personalized recommender systems for social networking. *Artificial Intelligence Review*, *37*(2), 119–132.

Zhu, Z., Zhou, Y., Deng, X. & Wang, X. (2019). A graph-oriented model for hierarchical user interest in precision social marketing. *Electronic Commerce Research and Applications*, *35*, 100845.

Appendix A

Glossary

User profiling/taste profiling - The process of pursuing user representative knowledge from such short text microblogs for better provision of third-party related services.

User interests profile - A data structure representative of a group's/individual's degree of interest in a set of topics.

Online Social Networks (OSNs) - Web-based applications and related services that (i) explicitly allow for the construction of a public or semi-public profile within their system (ii) list other users with existing or potential relations with the original user (iii) allow the individuals to navigate their list of connections. The connection types vary per the OSNs.

Vector - a tuple of data values. For example, a NumPy array in Python would be one. Basically, its how terms in (word2vec) and characters in (FastText) are represented.

Concepts - building blocks of a way in which a sentence or tweet is perceived.

Topic - A representation of a collection of related terms that are semantically relevant in a document.

Clustering - Process of grouping a set of objects or data points in a way that similar

ones are in the same group (cluster).

Cluster - A set of similar objects.

Topic Modelling - Text mining approach in the discovery of hidden semantic structures e.g. word and phrase patterns in a text body.

Labelled data - Data with meaningful classes or labels annotated for every observation or row in the dataset. Applicable in supervised learning.

Unlabelled data - Data applicable in unsupervised learning scenarios where meaningful tags or labels are not necessary.

Bag-of-Words(BoW) - A representation of text that describes the occurrence of words within a document with word count as a feature.

Cosine Similarity - Metric used to measure the similarity of documents where the cosine of the angle between two document vectors is projected in a multi-dimensional space. The smaller the angle, the more similar the documents and vice versa.

Tweet-metamodel matrix - A representation of the semantic relevance of a tweet to the metamodel topics.

Follow-backs - A representation of a bi/uni-directional relationship between users on short text microblogs that is determined by their shared interests.

Low-dimensional vectors - Vectors with typically 50–500 dimensions.

Word Embeddings/Vectorisation - Methodology in NLP to map words or phrases from vocabulary to a corresponding vector of real numbers which used to find word predictions, word similarities/semantics.

Silhouette Coefficient - Measure of the similarity between objects to its own cluster (cohesion) and compared to other clusters (separation).

Tweeters - Users on Twitter platform.

Text Mining - process of extracting information from structured sources.

Unsupervised learning - Knowledge discovery mostly applied to unlabelled data.

Semi-supervised learning - Knowledge discovery in both labelled and unlabelled data.

User interests - Specific aspects important in the identification of user(s) activities and in the perspective on short text microblogs.

User model - Data structure that is representative of a user's characteristics.

User profile - Actual representation of the user representative model

Weighting scheme - Function that determines the importance of the user interests.

Knowledge graphs - Knowledge bases with an ontology.

Metamodel - A set of topic label vectors derived from long texts to guide the learning process in shorter texts.

Topic - Collection of words or phrases that refer to a popular but temporal concept.

Tokens - List of unique words in a sentence/document/tweet for user input in a modelling algorithm.

Centroid map - List of terms and respective cluster labels.

FMI score - Geometric mean of the pairwise precision and recall between the true and predicted labels.

Homophily - Tendency of users to have positive ties with other similar users in socially significant ways.

Responsibility matrix - Individual users' levels of interest in topical clusters. Represented as cosine distance measures between documents/tweets and the topical clusters of interest.

Appendix B

Figures

B.1 Topical Term Distributions

The graphs below represent sample quarterly word counts verses term weights in topical interests across 2018, 2019 and 2020. Word counts and related weights in the identified topics are pertinent in most topical modelling algorithms.

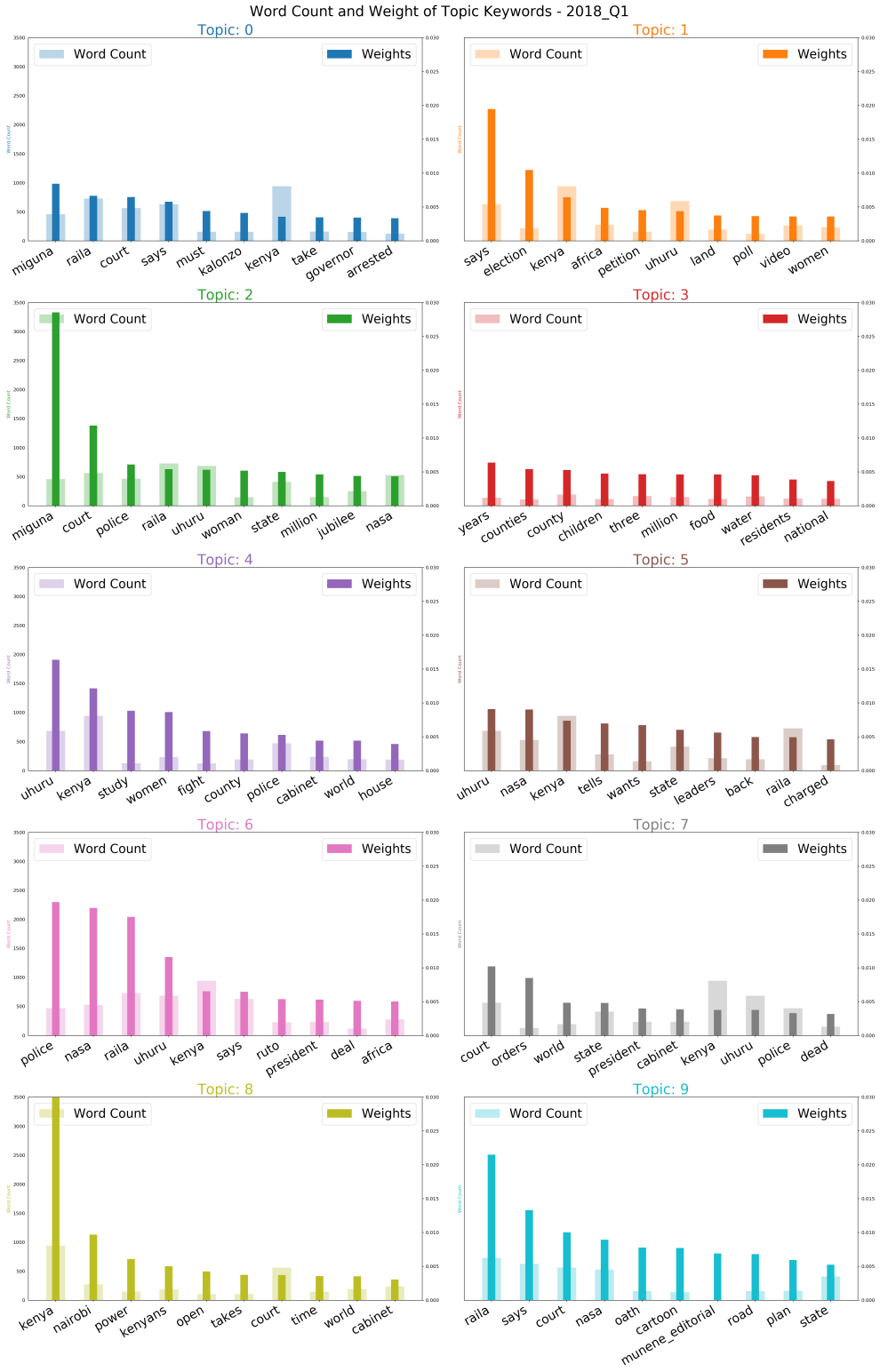


Figure B.1: Sample word count verses term weights in each topic in Q1 of 2018

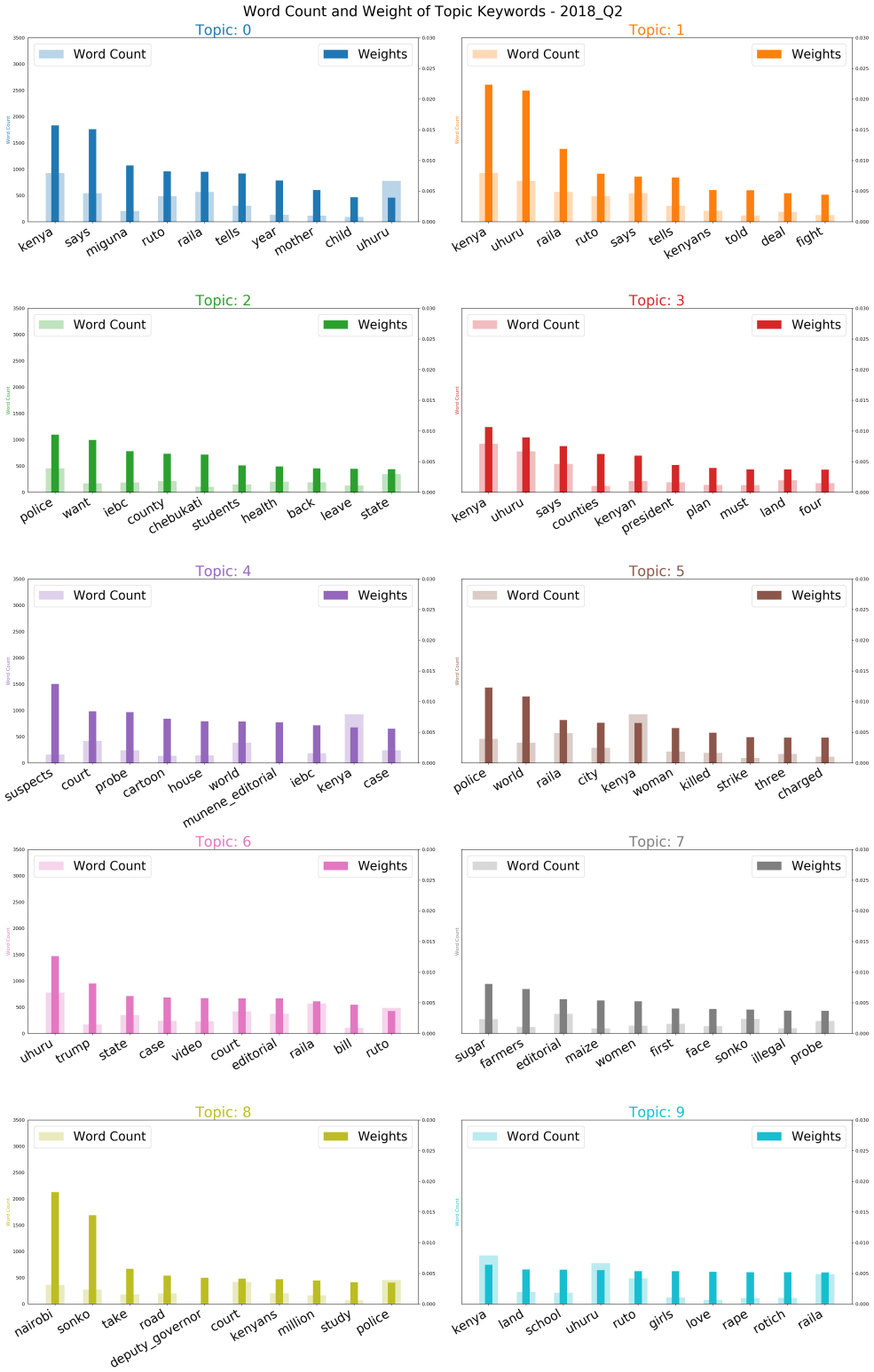


Figure B.2: Sample word count verses term weights in each topic in Q2 of 2018

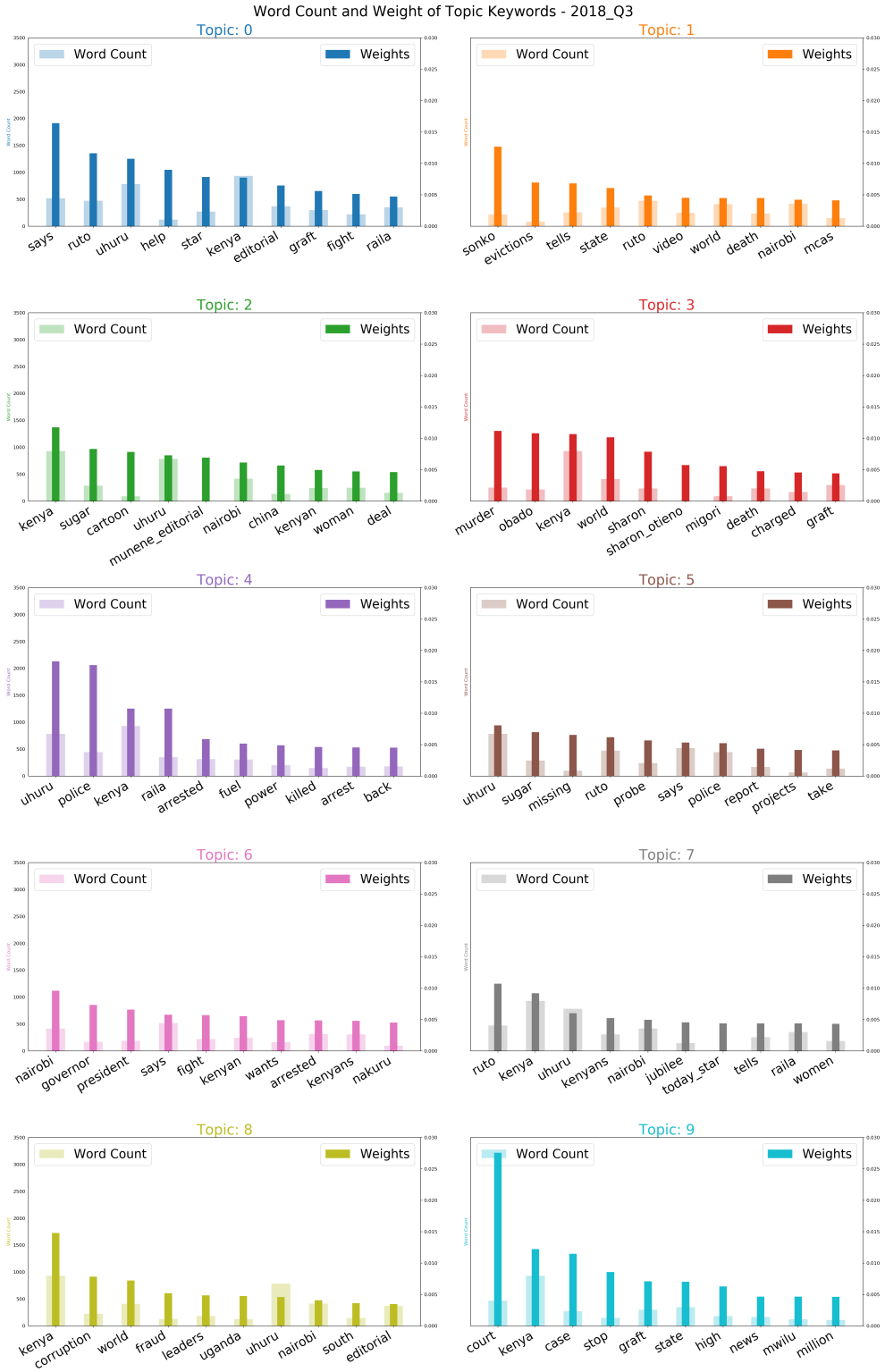


Figure B.3: Sample word count verses term weights in each topic in Q3 of 2018

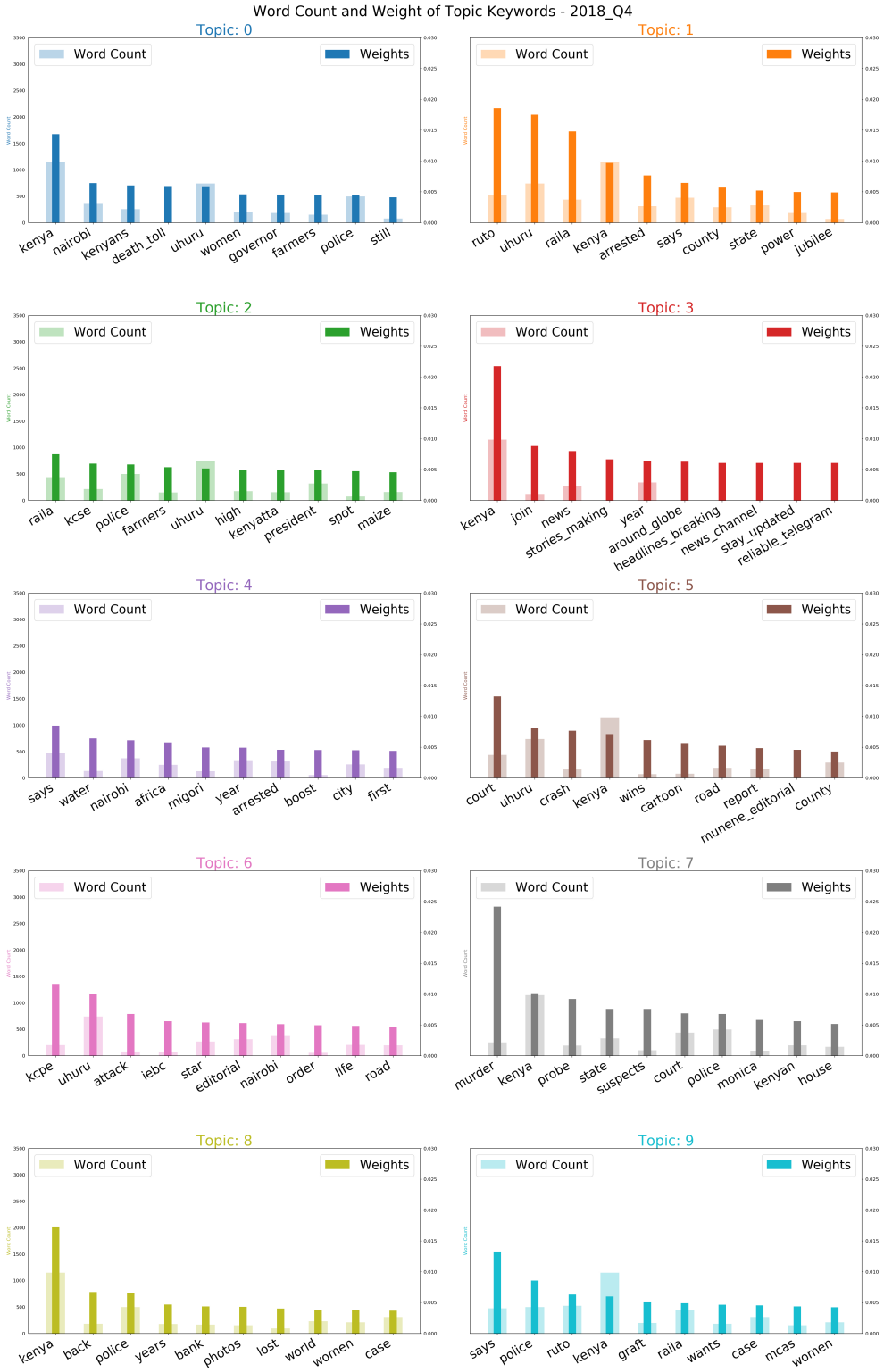


Figure B.4: Sample word count verses term weights in each topic in Q4 of 2018

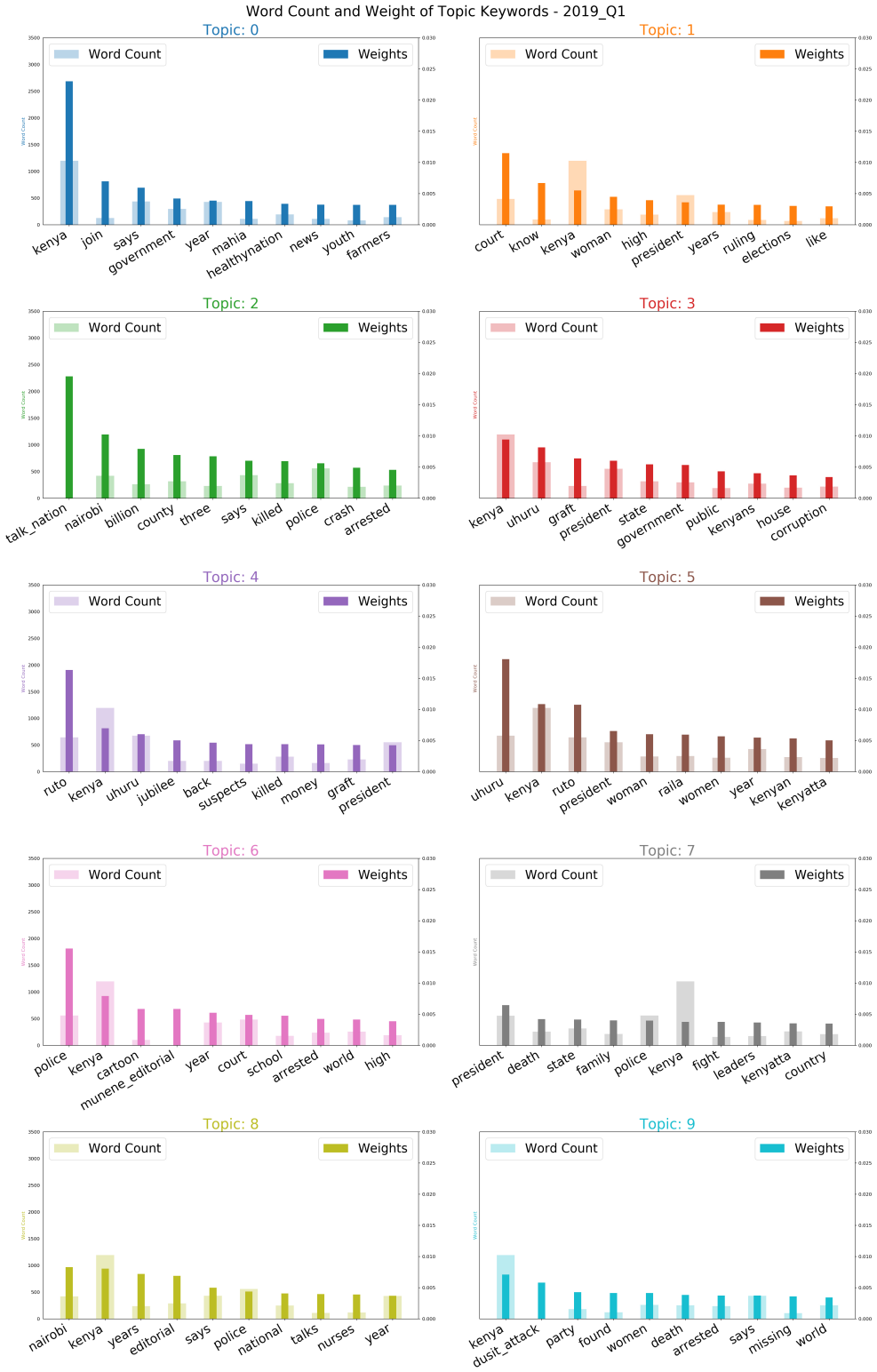


Figure B.5: Sample word count verses term weights in each topic in Q1 of 2019

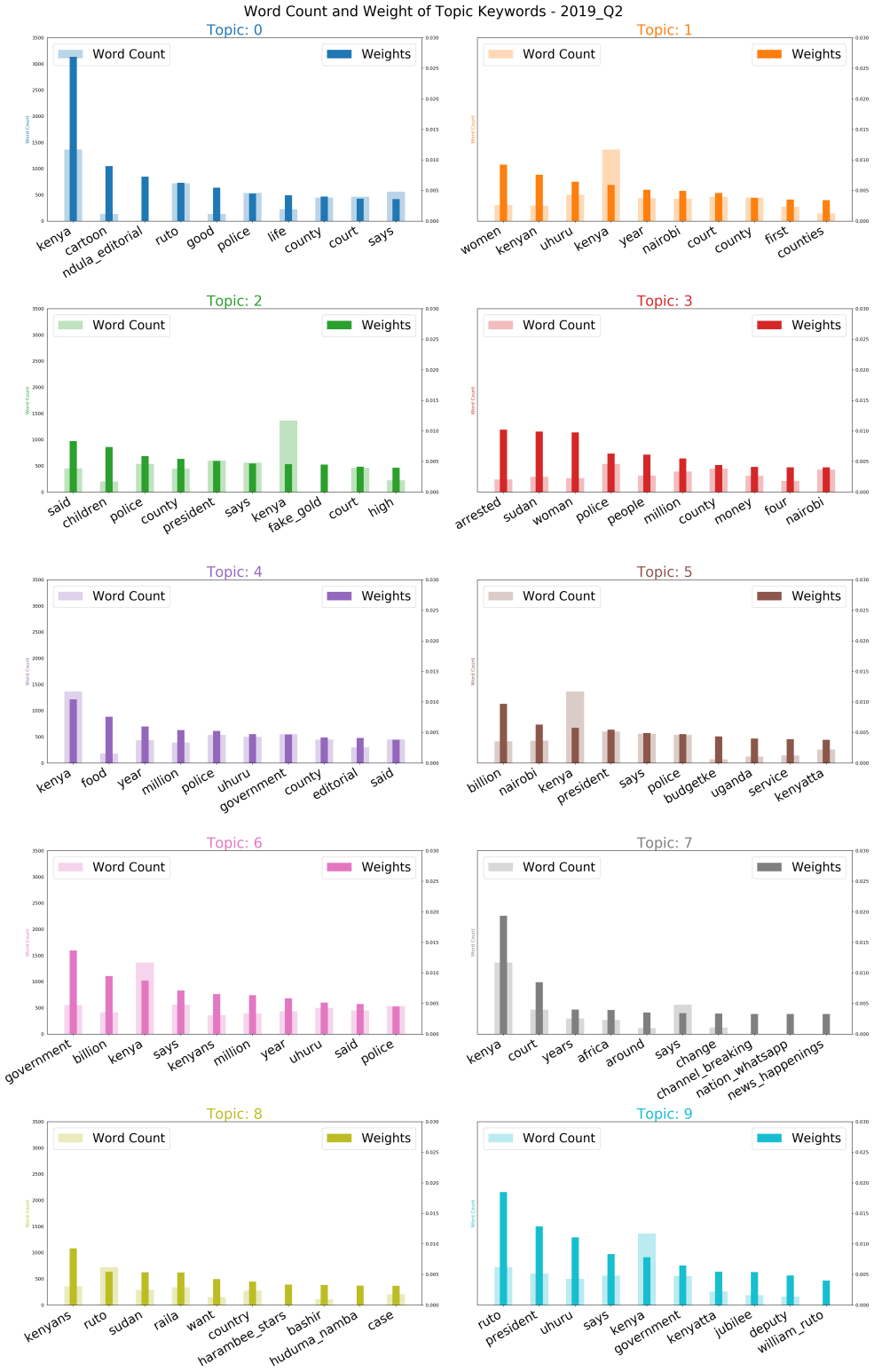


Figure B.6: Sample word count verses term weights in each topic in Q2 of 2019

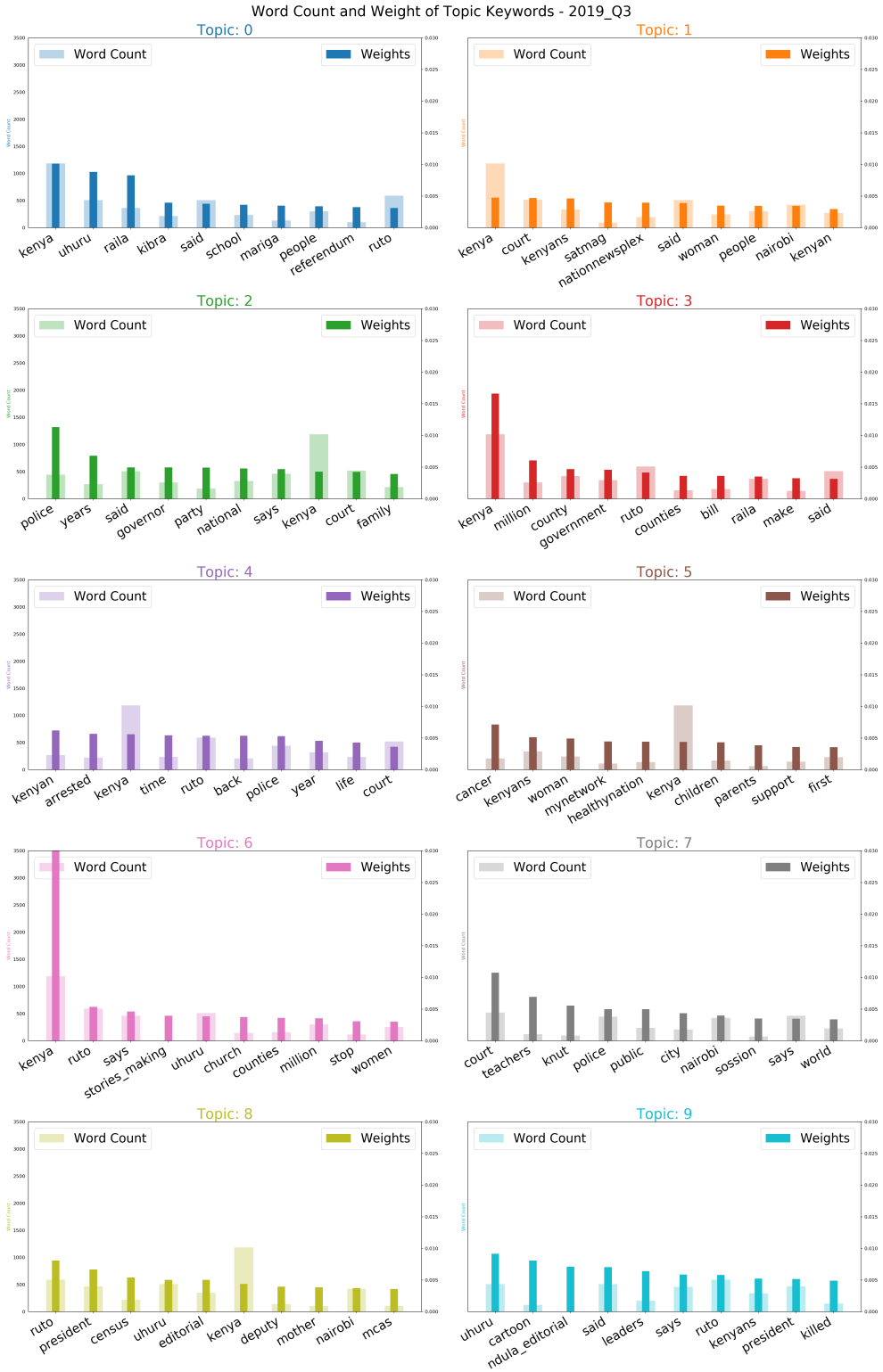


Figure B.7: Sample word count verses term weights in each topic in Q3 of 2019

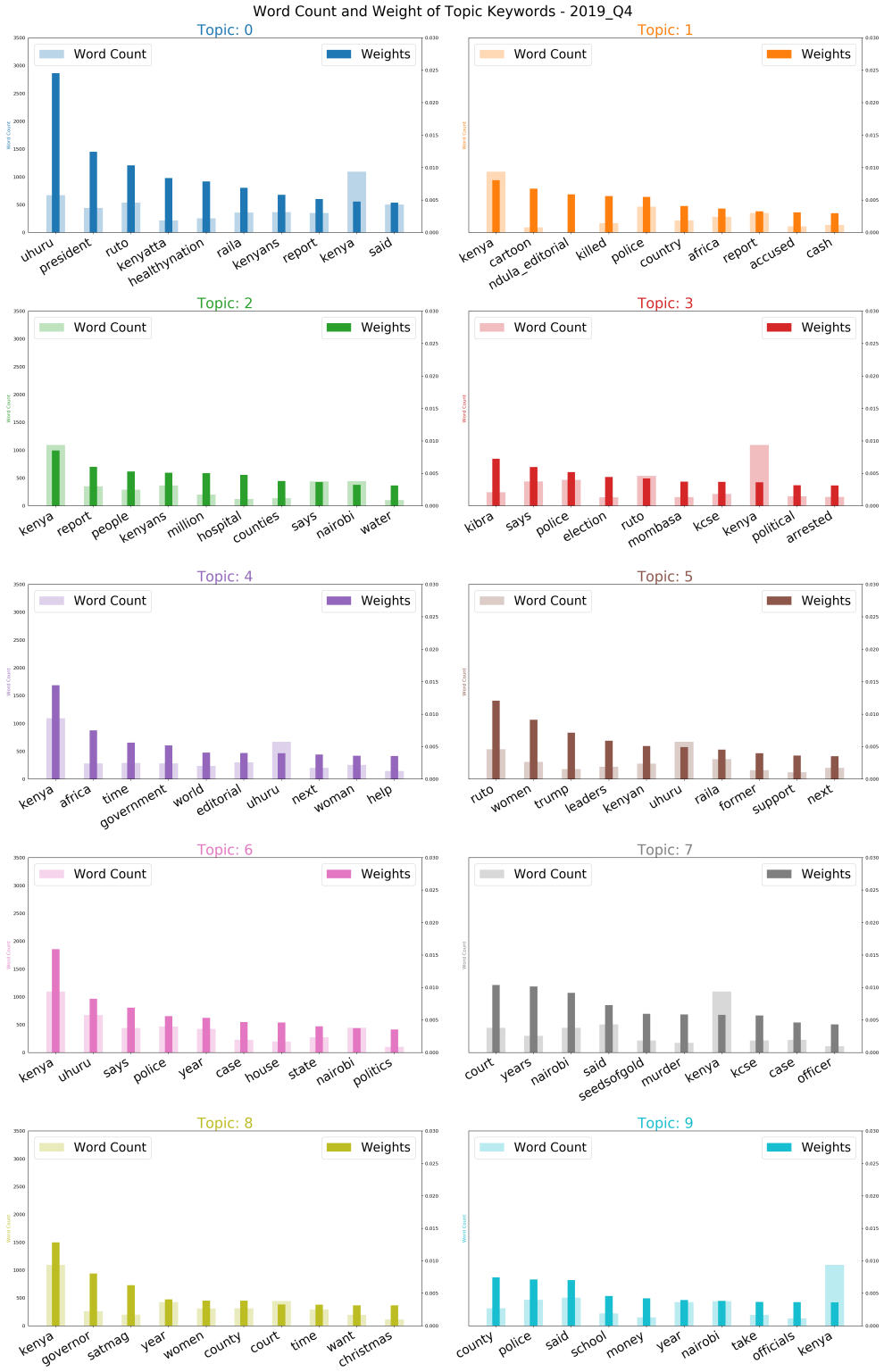


Figure B.8: Sample word count verses term weights in each topic in Q4 of 2019

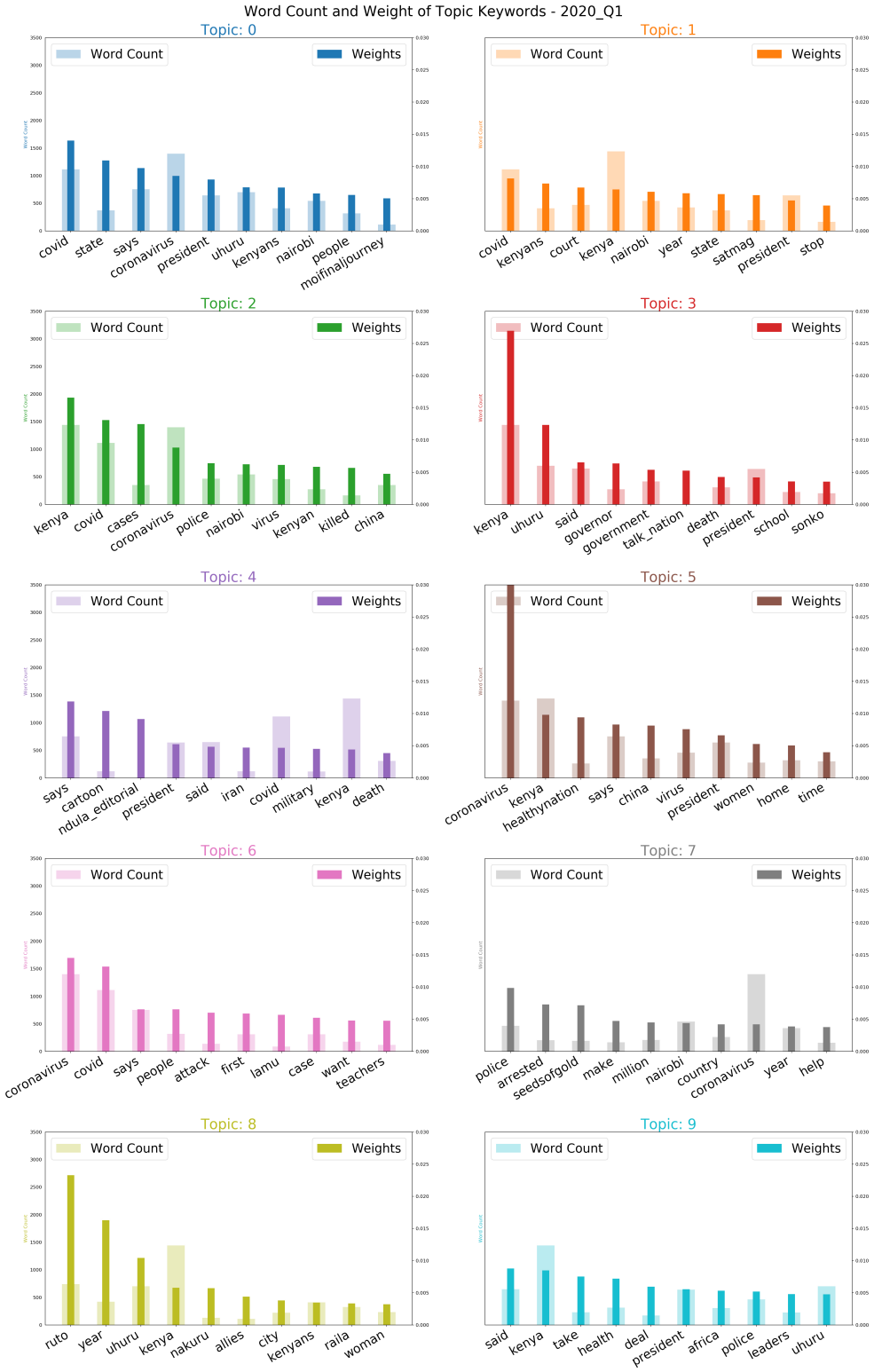


Figure B.9: Sample word count verses term weights in each topic in Q1 of 2020

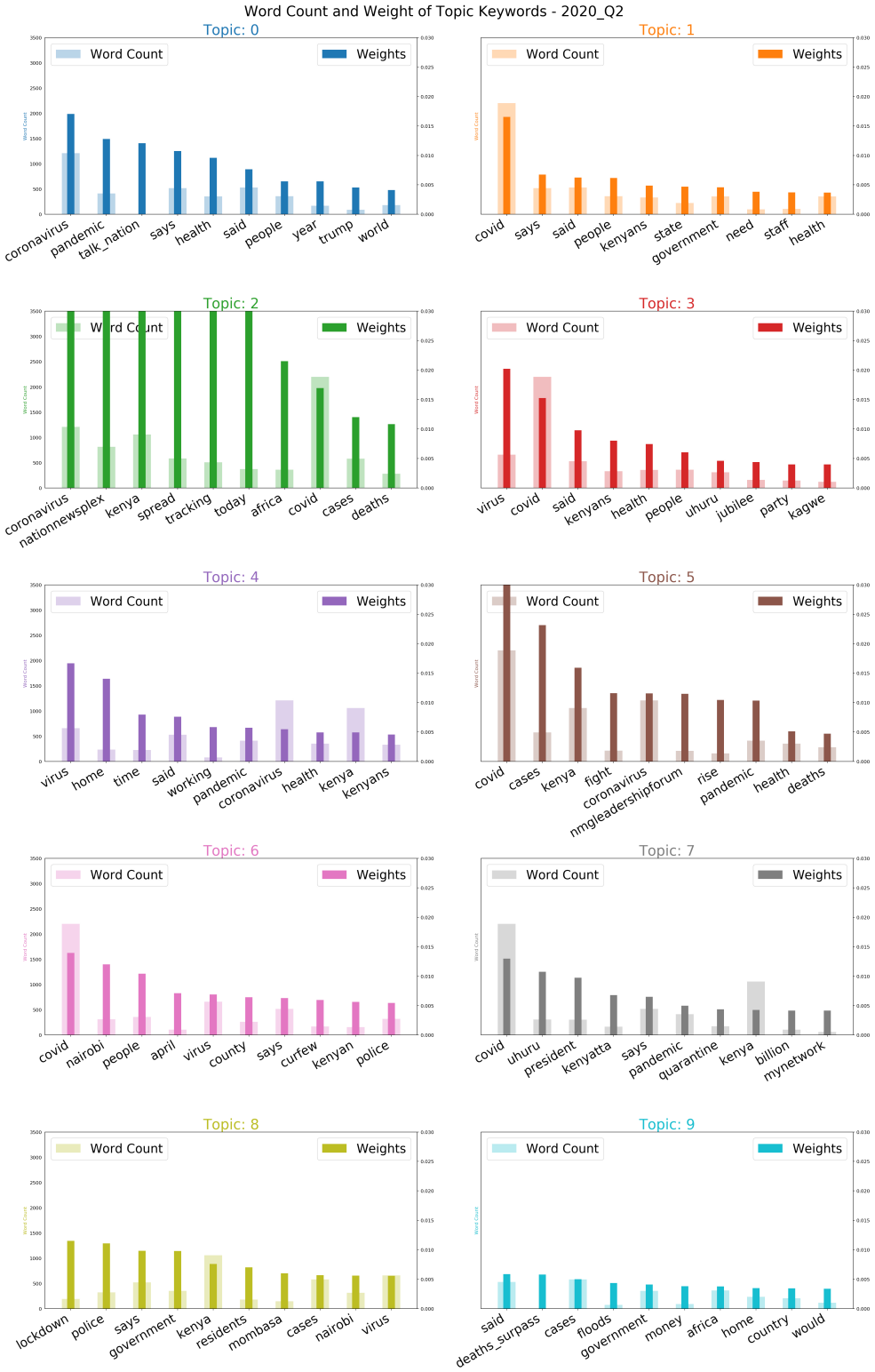


Figure B.10: Sample word count verses term weights in each topic in Q2 of 2020