AN INVESTIGATION OF FALSE INFORMATION AND ITS DETECTION

A THESIS SUBMITTED TO AUCKLAND UNIVERSITY OF TECHNOLOGY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF COMPUTER AND INFORMATION SCIENCES

Supervisor

Dr Ji Ruan

June 2023

By

Tatsuki Hashimoto

School of Engineering, Computer and Mathematical Sciences

Abstract

This thesis investigates the facets of false information and its detection mechanisms. Artificial Intelligence (AI) is currently employed in diverse tasks, such as social media recommendation, object detection, and speech recognition. However, it also presents challenges related to the spread of personal and misleading information. Such challenges encompass the manipulation of opinions, the birth of rumors, scaremongering, and eroding trust in governments. The methods to combat false information include fact-checking sites (which involve manual processes), automated fact-checking, and user awareness. Large Language Models (LLMs) are neural networks with over a million parameters that are pre-trained on large text datasets and fine-tuned on specific tasks such as question answering, language translation and sentiment analysis. LLMs are useful because they can be fine-tuned to classify and generate for specific tasks without requiring large datasets. We make two kinds of contributions. On the technical level, we have enhanced the accuracy of misinformation detection on a Twitter dataset using LLMs and probed the explainability of such algorithms through Explainable AI frameworks. On the knowledge level, we further investigated the societal aspects of false information issues. We examined the role of social media algorithms in amplifying false information and explored the potential of laws and regulations in addressing and mitigating the associated challenges.

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

I would like to express my gratitude to my supervisor Dr Ji Ruan for the support, guidance and knowledge on this thesis. I would like to thank my parents for supporting me in continuing to study in this subject. Finally, I would like to thank Auckland University of Technology for giving me the opportunity to study and complete my masters.

Chapter 1

Introduction

Social media companies are using AI to generate more revenue by keeping users on their platforms longer. As more users keep using their products, more information is gathered from the users, thus improving the algorithm's performance and gaining profit (Wong, 2019; Cobbe, 2020). The most popular social media platforms (Facebook, Instagram, TikTok and Twitter) are all driven by AI algorithms to give personalised feeds to each user. Most of these algorithms' inner workings are private, but there is enough information on parts of the algorithms to show how much personal data is collected. There is a risk of spreading false information via biases from recommendation algorithms (Cinelli, Morales, Galeazzi, Quattrociocchi & Starnini, 2021; Chen, Pacheco, Yang & Menczer, 2021). Thus in research on fake news from Twitter, Kumar and Shah (2018) found that false cascades (broadcasting) are significantly more profound than true cascades (peer-to-peer) and far more likely to spread faster. Novel information tends to be retweeted, but one cannot claim it is the only cause. False information is spread more because of human judgement than bots (Kumar & Shah, 2018). Bots are also a part of the problem, as they are part of a network of bots used by third parties to post or retweet false information to influence, gain profit, and engage in malicious activity (Kumar & Shah, 2018). With the spread of false information, users are going to

fact-checking websites, which is one of the most popular ways to inform themselves whether the information they see on social media is false or true. Manual fact-checking is currently a time-consuming process, and this is where LLMs come in. With the rise of data generation, sharing, and AI use, there is an increase in personal data stored and analysed by tech and social companies. There are data protection acts, but are they enough to protect personal data, and are the proposed AI laws enough to protect us from data use and the spread of false information?

Research Questions In this thesis, the goal is to investigate false information and its detection, the effects of social media algorithms, and data regulations. These three parts are related to each other, and we wanted to mitigate the negative impact of false information on society. Specifically, we address three main research questions.

The first question is on how to detect false information. This question includes false information, its impact, determining whether the information we see is false or true, and the methods to detect false information. There has been much research on false information in recent years, and the rise of false information, especially in the 2016 US presidential election and the COVID-19 pandemic, has had a concerning effect on society. False information detection is a relatively new area for AI research, and several systems have been used to combat false information, including Claimbuster (Hassan et al., 2017) and TOKOFOU (Tziafas, Kogkalidis & Caselli, 2021). The objective is to improve existing methods and provide a level of explainability to a method.

The second question is what are the insights of social recommendation systems on the most popular social media platforms and their effects on society. The fine details of social recommendation algorithms are usually hidden from the public, and users usually do not know how much personal information is gathered and processed. The objective is to analyse the different algorithms/systems used to recommend users' content and the issues/dangers of its user's information.

The final question is how data and AI regulations can affect users and technology companies. This question includes data protection laws, its issues, and personal data and its protection. With the increase in AI and data use online, there is a huge increase in personal information being stored and analysed by companies and governments globally. The objective is to explore current and upcoming laws on personal data and AI regulations. In addition, we analyse big technology corporations and their privacy features.

Study Contribution In this thesis, we make two kinds of contributions:

1. On the technical level, we improved the multi-lingual capability of the TOKOFOU false information detection system. We compared it with the latest Generative Pre-trained Transformer 3 (GPT-3) and GPT-3.5 (ChatGPT) to determine whether a tweet is false information. We also explored explainability of our detection algorithms using LIME (Ribeiro, Singh & Guestrin, 2016) and SHAP (Lundberg & Lee, 2017) framework.

2. On the knowledge level, we further investigated the societal aspects of false information issues. We examined the role of social media algorithms in amplifying false information and explored the potential of laws and regulations in addressing and mitigating the associated challenges.

Thesis Structure Chapter 2 is the background on false information and its detection methods. Chapter 3 presents our work on false information detection in a task that originated from the 2021 workshop on fighting the COVID-19 infodemic. We developed TOKOFOU_T, which improved an existing algorithm (TOKOFOU) on its multilingual capability. In addition, we managed to use the very recent GPT models (GPT3Ada with fine-tuning and chatGPT3.5 Turbo without fine-turning) on the same task. Chapter 4

discusses explainability using two common explainable models (LIME and SHAP). LIME and SHAP are used to explain the output of TOKOFOU in some tweets and are compared to each other to find common keyword similarities. Additionally, LIME is used to explain the fine-tuned GPT3 Ada. Chapter 5 discusses the algorithms used by social media companies and their comparison. In addition, we explored the effects of the recommendation system on society. Chapter 6 discusses the regulation of data and AI, including privacy, content laws, surveillance capitalism, and the policies of technology companies on personal data, such as Apple and Google. Chapter 7 concludes the thesis, including limitations and future work.

Chapter 2

Background on False Information

In this chapter, we will present the findings of false information, its impact and methods to detect them. There have been several works on false information, from the type of false information and how it affects society to the methods to detect it manually and automatically.

2.1 False Information and its Impact

False information is categorised into misinformation and disinformation (distinguishing them by their intent and content to their victims). False information can have a destructive impact on our daily lives and change the outcome of an election (the presidential election in the USA). Humans have poor judgement in identifying a range of false information, whether trained or causal (Kumar & Shah, 2018). Moreover, the distraction of social media (paying no attention to the source) can enable the readers to be tricked (van der Linden, 2022). According to Kumar and Shah (2018), the more compelling the content is (e.g., referenced, written well, and long), the less we can identify it from false or true information.

2.1.1 Types of False Information

Misinformation

Misinformation is a type of false information that tends not to deceive the victim but to misrepresent content or mislead victims, e.g., when users share posts from someone they know (Kumar & Shah, 2018; Ghanem, Rosso & Rangel, 2020).

- Rumours are stories that are not confirmed, spread through people, and can damage reputations (Cho, Rager, O'Donovan, Adali & Horne, 2019; Vosoughi, Roy & Aral, 2018). Depending on the intent, it can be misinformation or disinformation. An example of this recently is "Did Betty White say she got COVID Booster 3 Days Before She Died"[1] which is false and scared people not to take the COVID booster shot.

- Fake/False News are new stories that look like real news and spread unreliable information (Vosoughi et al., 2018). These include satire, parody, fabrication, and photo manipulation (Karduni, 2019).

- Satire is content that uses humour, exaggeration, and irony. Although disclosed, they can be shared on social media without disclosing as satire. This can be seen in TheOnion content on social media (Zannettou, Sirivianos, Blackburn & Kourtellis, 2019).

Disinformation

- Propaganda are stories to mislead and harm the victim in a political context. It can lead to a change in history using visuals or sounds such as banners, music, and posters (Kumar & Shah, 2018).

[1] https://www.snopes.com/fact-check/betty-white-covid-vaccine-booster/

- Clickbait uses misleading headlines/thumbnails to get victims to open the article to gather more views or take it out of context. Although this is disinformation, it is the least severe type (Kumar & Shah, 2018).

2.1.2 Social Impact

False information affects our society, including not trusting the government, confusion, and changing or swapping people's opinions (Criado, Sandoval-Almazan & Gil-Garcia, 2013; Skurnik, Yoon, Park & Schwarz, 2005; van der Linden, 2022). Events such as natural disasters and terrorist attacks produce engagements, which increases false information (Kumar & Shah, 2018).

Criado et al. (2013) found studies that show that having governmental officials and services on Twitter can improve the response to false information but can have the opposite effect if they are not supervised (as seen in Trump's Twitter).

De keersmaecker and Roets (2017) found that people would have a lower response to having learned that their answer to whether the information was true was false when they have lower cognitive ability. Thus, their attitude changes less than that of higher-cognitive people. This study concluded that false information can still have a lasting effect even if the victim is told the true information, which depends on cognitive ability.

False claims can be misremembered as true when readers repeatedly see the same claim because of the familiarity (van der Linden, 2022). This is seen in the study in Skurnik et al. (2005), where older individuals would more likely misremember false information as true information when told three days later. Similar false information can make individuals think they have already seen the information before, believing it to be true. Results of the two experiment study showed that older individuals are more at risk of claiming false information as true information than younger individuals. "false memory not only depends upon the credibility of the source but also depends upon the

credibility of the medium of presentation" (Fenn, Griffin, Uitvlugt & Ravizza, 2014, p. 1555). This result shows individual remembers from the medium (participants were shown three images, two from Twitter and another type as seen in Figure 2.1) (Fenn et al., 2014)

Figure 2.1: Tweets types
From *The effect of Twitter exposure on false memory formation* (Fenn et al., 2014, p. 1553)

A study on hoaxes on Wikipedia shows that hoaxes last longer with fewer viewers but gain more views as it stays up. This type is impactful as many individuals rely on/credit Wikipedia, spreading further to other sites (Kumar & Shah, 2018). Web Brigades (Russian trolls/bots) and Click farms (groups of workers to click on advertisements) attack Wikipedia by hijacking vulnerable wiki pages, such as wikis that do not have enough information on a particular topic, political topic, or religious topic (Saez-Trumper, 2019).

False information in advertising is impactful because of its nature to deceive customers. Nuseir (2018) found that customers will lose trust when they encounter false advertisements for a well-known company/service. Thus, a company with a high market share advertises truthfully. In 2016, Facebook banned false news advertisements via links after concerns about false news about the US presidential election (Chiou & Tucker, 2018). The ban had a 75% decrease in shared false news compared to Twitter, which did not have a ban. Laws are put into place to stop false news, some examples

are [2]:

- Political ads to be transparent (include authors and funders) in Australia.
- Bangladesh introduced a bill to imprison "propaganda", "aggressive and frightening content " spreaders.
- The United States introduced a bill requiring social media to keep records of ads and publicly pay attention to who is paying.

A Facebook group is one way to spread false information via an echo chamber. Chiou and Tucker (2018) found that anti-vaccine groups exist to share information instead of creating events and sharing activities. Most shared information links to "government fact-based" sites but have misleading titles or conclusions. The posts are published by several users in the group (1% of users) and shared by the members to other groups, thus spreading it further (Chiou & Tucker, 2018). Schmidt, Salomon, Elsweiler and Wolff (2021) surveyed Facebook users, and the results showed that they occasionally review the post's comments but rarely check whether it is liked/shared by trusted sources. Users sometimes unsubscribe from the publisher if they see false information and rarely comment/share the post (Schmidt et al., 2021).

2.2 Fact Checking Information

Traditionally, fact-checking is a process where journalists hire fact-checkers to verify and proofread claims in their articles. Modern fact-checking[3] consists of dedicated websites to verify claims in social media. Two events raised fact-checking awareness and sites, such as the 2009 Pulizer prize, which launched PolitiFacts, 'fake news'

[2]https://www.poynter.org/ifcn/anti-misinformation-actions

[3]https://www.poynter.org/fact-checking/2016/who-decides-whats-true-in-politics-a-history-of-the-rise-of-political-fact-checking/

and recently 'COVID-19' (Ireton & Posetti, 2018; Seaton, Sippitt & Worthy, 2020). Fact-checkers help us understand misinformation, whether mistaken or intentional, by explaining the reasons for the claim.

2.2.1 Fact Checking Websites

One easy way to check whether the information is false is to check fact checking websites such as well known fact checking sites are Factcheck.org, Snopes, PolitiFact, and Fullfact (Nakov et al., 2021). These websites are checked by creditable/primary sources[4]. An example of a source is the Library of Congress, the Congressional Budget Office, the Congressional Research Service, the Kaiser Family Foundation on health care data, and experts on a particular topic (Karduni, 2019). Fact checking sites[5] select topics from the greatest user requests or searches, such as trending social media topics, and submit them on Twitter or Facebook. Some Fact checker sites[6] work with Facebook to debunk links, images, and videos that users or systems have flagged in Facebook posts. Google[7] has a tool that users can use to search for fact check topics. This is similar to the regular Google search, but only for searches with publisher submitted information.

2.2.2 Automated Fact Checking with AI

Due to the rise of social media and news, manual fact-checking can not keep up with the amount of information being created/shared. Therefore, automated fact-checking is important for detecting false information. Nakov et al. (2021) lists several needs of the fact checkers, such as methods to find worthy claims (ranking or prioritising claims), detection of already debunked claims (specifically complex claims), retrieval of evidence

[4]https://www.factcheck.org/our-process/, https://www.snopes.com/transparency/
[5]https://www.snopes.com/about/
[6]Factcheck.org
[7]https://toolbox.google.com/factcheck/about

(images and videos) and auto verification (for complex claims and reliability). False information content, such as images, links, and text, provides fact-checkers and models that can help determine whether the post's content is false or true. One such AI model is called Claimbuster (Hassan et al., 2017). This model uses Natural Language Processing (NLP) and supervised learning to generate scores to help check for false information in the publisher's content. Claimbuster was improved by using fact-checking websites (Google fact check tools) to find similar content (Karduni, 2019; Hassan et al., 2017). Another AI fact-checking algorithm is TOKOFOU, an LLMs-based model that answers seven questions on COVID-19 tweets (Tziafas et al., 2021).

ClaimBuster

Hassan et al. (2017) explains ClaimBuster, which features monitoring live debates, tweets, and news by using its repository and querying. This model was tested in the live 2016 U.S. election debates, where it would pick up sentences from the candidates and post a score based on their truthiness. ClaimBuster has five modules in the system. Claim Monitor is a module that collects and monitors various sources such as broadcast media (TV programmes), social media (Twitter), and websites. Once there is a new claim, it scans the content using the Claim Spotter module, which generates a score based on the sentence's factual, subjective, and opinion levels (a lower score is more opinionated and subjective). High-scoring sentences from the Claim Spotter are passed to the claim matcher, where it searches for related claims via fact-checking websites and its repository. If there is no matching claim, the claim is passed to the claim checker. The claim checker generates questions based on the claim, which are searched in Google and the Wolframe API. Once the claim is matched with the results, it is sent to the Fact-Check Reporter. This module combines the two results from their social platforms and repository with the claim checker score.

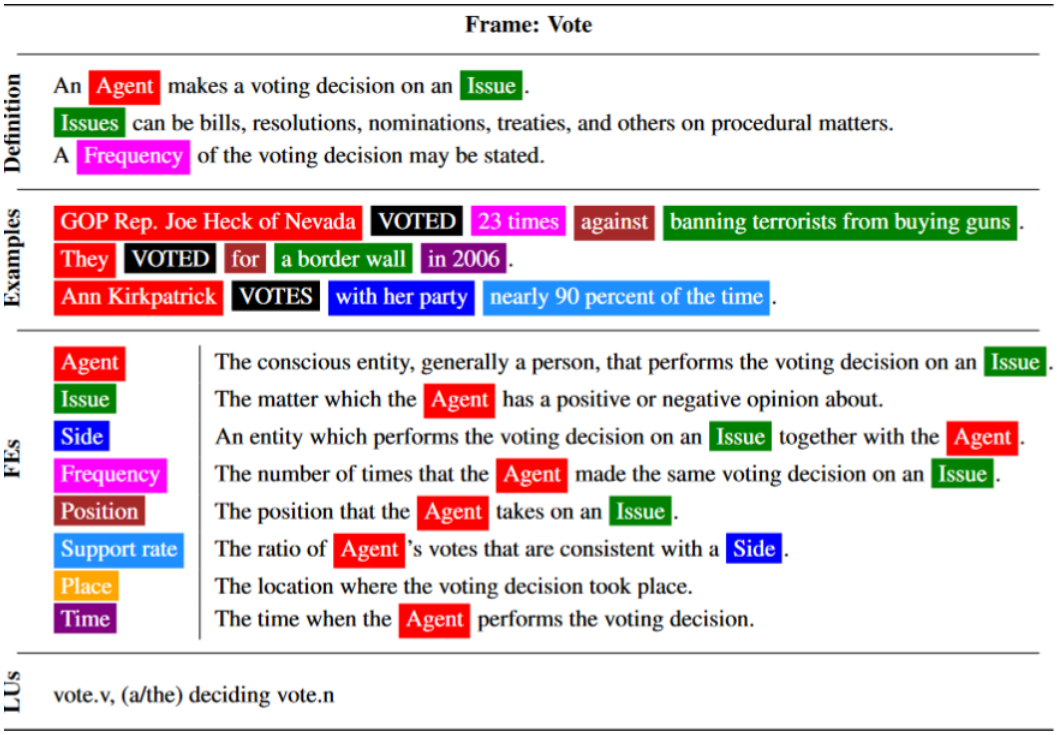


Figure 2.2: The vote frame

From *Modeling Factual Claims with Semantic Frames* (Arslan, Caraballo, Jimenez & Li, 2020, p. 2512)

Later articles of ClaimBuster researched improving claim spotting and claim matching using semantic frames to model factual claims. Arslan et al. (2020) introduced an extension of FrameNet, a theory that humans "understand things by performing mental operations on what they already know" (Ridcully, 2003, p. 3). This knowledge creates a structure to describe events and relationships (Ridcully, 2003). A frame consists of rules[8]: Examples, FEs (Frame Elements) and LUs (Lexical Units), which enable sentences to be annotated. Figure 2.2 shows a vote frame that the author has made based on PolitiFact claims and structures (Arslan et al., 2020). Semantic frames allow them to find patterns and group the claims into groups. Thus, improvements can be made in claim spotter and matcher. Furthermore, it would make manual fact-checking more efficient by identifying "statements of fact" (sentence annotation), claim duplication, and "translating claims to structured queries" (Arslan et al., 2020, p. 2518) to be

[8]https://www.nltk.org/howto/framenet.html

verified by approved sources. This reduces the time it takes to research and confirm the claim (Arslan et al., 2020).

TOKOFOU

Tziafas et al. (2021) introduced TOKOFOU, an algorithm that answers seven questions from a COVID-19-related tweet and was created for the NLP for Internet Freedom 2021 (NLP4IF) workshop. The seven questions are:

1. Is it verifiable factual claim?
2. Is it false information?
3. Interest to general public?
4. Harmfulness?
5. Need verification of claim?
6. Harmful to society?
7. Requires attention to government?

Further information can be seen in Table 3.7. TOKOFOU allows us to determine whether the tweet contains false information based on the seven questions. TOKOFOU uses three models and four fine-tuned models for specific questions. The outcome is determined by majority voting (further detail in Figure 3.5). TOKOFOU ranked top in the NLP4IF 2021 workshop in detecting COVID-19 for the English dataset. TOKOFOU can be an aid/assistant for fact-checkers or an "annotation of Twitter data for misinformation" (Tziafas et al., 2021, p. 122)

2.2.3 User Awareness

Bringing awareness to misinformation is a crucial skill to learn to combat misinformation. One such way is by getting internet companies to show ways to detect and handle misinformation. Verizon[9] provides a guide to spot and combat false/fake news, such as what is misinformation, misinformation vs disinformation, types of misinformation and examples, recognising fake news and misinformation, and how to handle/report them on social media. Teachers[10] can bring awareness of misinformation to their pupils by creating posters and quizzes, thus teaching them to handle misinformation. These skills are essential because a survey of Facebook users showed that they disregard sources, spread posts if they are interested in a specific topic, and are reluctant to utilise tools to help handle false information (Karduni, 2019).

Chang et al. (2020) surveyed a card game called LAMBOOZLED!, young students play this game to identify misinformation. The skills they learn from this game is :

- Using evidence to handle bias.
- Looking at / questioning content to find features (e.g., suspicious URL, modified images).
- Fact-checking skills (collecting evidence, looking for sources).

The game was played either in playtesting workshops or classrooms. They found that it reveals that the experience earned in the game can be applied well in the real world, and students are also engaged/participate. Teachers who are well prepared, e.g., show a video of the game and prepare lessons and resources, allow the students to play the game easier, but on the other hand, when there is no preparation, students are lost and do

[9] https://www.verizon.com/info/technology/fake-news-on-social-media/

[10] https://resourced.prometheanworld.com/fighting-fake-news-modern-classroom/, https://www.who.int/news/item/27-07-2021-raising-awareness-of-misinformation-among-children-in-poland

not engage with the game. Therefore, students would not benefit from playing the game if the teachers did not explain the game adequately. Hodgin and Kahne (2018) surveyed teachers who support students in handling misinformation at school and suggested three approaches:

- Develop Nuanced Skills & Strategies: learning skills beyond looking at rules and using a checklist by students doing analysis of sources and questioning the claim.
- Reflect on Though Processes: teaching students that their opinions and bias would affect their evaluation of the claim and reflect on their judgement.
- Practice, practice, practice: students practice their learned skills through weekly event topics or apply them to through school curriculum.

Having a routine allows students to apply their skills and be confident in researching and reading claims. The approach had success with teachers they collaborated with using media learning (Hodgin & Kahne, 2018).

2.3 Methods to Detect False Information

In this section, we explore several methods to detect false information including content-based, graph-based, social context-based, modelling-based and other methods.

The findings in van der Linden (2022) show that the best practice to debunk misinformation is by first giving victims easy and simple facts using sources, mentioning the myth surrounding the claim once, explaining how and why the claim is false, and finally ending with facts and alternative sources. These practices show better effectiveness than other methods.

2.3.1 Content-based Method

Content-based method uses the content of the post to determine whether it is false information or not. One of the post's main content that models use is text, which contains valuable information such as lexical features, syntactic features, topic features, and semantic features (Guo, Ding, Yao, Liang & Yu, 2020). Furthermore, URL links, hashtags, and images can also be used to classify false information. AI, such as Support Vector Machine (SVM) and decision tree classifiers, is used for content-based detection. Two such examples of content-based method are ClaimBuster (Hassan et al., 2017) and TOKOFOU (Tziafas et al., 2021). Content feature detection is an important method to detect false information, as humans use content to interact and make decisions. Crowd intelligence detections use these features to collect social context (user interactions), collective knowledge (user evidence), and collective behaviours (groups of users' behaviours) (Guo et al., 2020).

2.3.2 Graph-based Method

Graph-based detection is an algorithm that uses the nodes of users to create a graph. This is commonly used for fake reviews, which show a connection between users and their reviews of different products (ratings, number of reviews, and ratio of singleton reviews). This algorithm is not a critical technique because of pretending users that camouflage amongst real users (Kumar & Shah, 2018). Another graph-related model is Graph Convolutional Network (GCN). This model is an extension of the Convolutional Neural Network (CNN) that uses graph inputs that contain feature matrices (nodes and features). Each layer is a feature that generates a feature representation (Dong, Zheng, Quoc Viet Hung, Su & Li, 2019; Guo et al., 2020). This can be applied to propagation claims using social profiles and tweet contents (Dong et al., 2019).

2.3.3 Social Context-based Method

Social context-based method uses social context (human interaction features) such as likes, comments, and tagging. These methods use neural networks (RNN, CNN) and Long Short-Term Memory (LSTM). Social context classifiers can find the connection between users and how the post can spread. This can be seen in a network graph or tree. By gathering emotional context from false information content, it can adequately categorise different types of false information by the emotions they display (Ghanem et al., 2020). Using emotional lexicon models, they discovered that clickbait displays more emotions than the others. The emotions can easily separate the types of false information into groups, although there is some overlapping of true information (Ghanem et al., 2020).

- Clickbait tends to display "surprised" and "negative emotions", e.g., "You Won't Believe what happened next"
- Hoaxes are difficult to classify as they do not have a pattern but seem to show "like" emotions e.g., "COVID-19 spreads through petrol pumps".
- Propaganda information shows "joy" and "fear", interestingly "calmness" e.g., "Don't let tobacco take you breath away".

(Ghanem et al., 2020)

2.3.4 Modelling-based Method

Modelling based method is a technique to create the spread of false information models and to prevent it. One example of this use is a propagation model created by (Acemoglu, Ozdaglar & ParandehGheibi, 2010). This model creates either normal or forceful nodes; forceful nodes change (some) of normal nodes' beliefs while forceful nodes change to some degree when they interact with each other. Thus "echo chambers" are formed.

Acemoglu et al. (2010) concluded that when several forceful agents update their beliefs from the information they obtained by individuals, they would have a greater spread of false information. Furthermore, members in small groups are influenced the same way as other groups by forceful agents that are not linked together.

2.3.5 Others

Guo et al. (2020) surveyed detection models, the early detection model is an important topic to be discussed as other models depend on situations that have already happened or similar occurrences/patterns. Deep learning is used for this topic to extract context, links, and hashtags to help classify the information. Another interesting model is to present reasons why the claim is false to better understand the claim. One way to show this is to visualise "attention degree" by using words such as post emotions, signal words, and textual features.

Zareie and Sakellariou (2021) recommends two strategies to minimise misinformation: one is to block it before readers read it (blocking-based), and two is to spread awareness of the false claim (clarification-based). Block-based methods have two types, node blocking (blocking nodes taking account of misinformation) and edge blocking (removing nodes by identifying critical edges). Clarification-based has two types, campaign-oriented and protection-oriented. The campaign-oriented method brings truth campaigns to nodes with the most influence from other nodes. In contrast, the protection-oriented method protects users from misinformation that is not affected by the campaign.

Chapter 3

Our Work on False Information Detection

3.1 Introduction

In this chapter, we will present our work on false information detection. We pick a task from the NLP for Internet Freedom (NLP4IF) 2021 workshop on fighting the COVID-19 Infodemic and Censorship Detection.

There are two tasks in the NLP4IF workshop. The objective of Task 1 is to design and create a system/method to answer (label) the seven questions (section 3.2) from a tweet from a given dataset in three languages (English, Arabic, and Bulgarian) to study "the problem from a holistic perspective" (Shaar et al., 2021, p. 82) (How do the seven questions provide the answer?). Task 2's objective is to predict what types of posts on Sina Weibo (a Chinese blogging platform) are removed and what is saved. Task 1 is the focus of our work.

The goal is to build an algorithm to answer the questions that humans would answer (ground truth) using a set of inputs and several outputs (7 questions), as shown in Table 3.1. There are seven labels on which the human and the algorithm use tweet_text to

answer the seven questions (description of the human to answer).

Table 3.2 shows the top 3 teams for the English track. Out of the four algorithms that do all three languages, advex (no published article) has the best performing system on average (3rd), as seen in Table 3.3 and 3.4. Compared to advex, HunterSpeechLab (rank average of fourth), InfoMiner (rank average of fourth) and spotlight (rank average of fifth).

We aim to improve the TOKOFOU system to label all three languages (English, Arabic, and Bulgarian), as TOKOFOU is the best performer for English but could not classify Arabic and Bulgarian well and did not participate in those languages. In addition, the top average rank in all three languages is **advex**.

Column Name	Detail
tweet_no	Tweet number
tweet_text	Twitter Tweet
q1_label	Verifiable Factual Claim - Yes if there is claim which is supported by evidence such as statistics, verifiable, factual information, examples or personal testimony. No if there is no evidence
q2_label	False Information - Yes if there is claim of false information
q3_label	Interest to General Public - Yes if the claim relates to potential cures, updates on number of cases, on measures taken by governments or discussing rumors and spreading conspiracy theories (Tziafas et al., 2021)
q4_label	Harmfulness, Yes if the claim is harmful to the society / person(s) / company(s) / product(s)
q5_label	Need of Verification, Yes if a fact checker is needed verify the claim
q6_label	Harmful to Society, Using the matrix in Table 3.6
q7_label	Require attention, Yes if the claim needs to be looked by government department.

Table 3.1: Training dataset columns
Note: Q2,Q3,Q4,Q5 would become nan if Q1 is no or "not sure"

TOKOFOU algorithm is a majority voting ensemble model that uses six pre-trained transformers. The pre-trained transformers are BERTweet (RoBERTa base model

trained on 850M tweets), CT-BERT (BERT large model trained on 160M COVID-19 related tweets), and TWEEETEVAL (RoBERTa base model trained on 60M tweets) using their fine-tuned models (hate-speech, emotion, irony, and offensive). Training "on 15 epochs on batches of 16 tweets" (Tziafas et al., 2021, p. 121) is done using the AdamV optimizer (learning rate of $3 \cdot 10^{-5}$ and weight decay of 0.01) on the provided English dataset without URLs (Tziafas et al., 2021).

Rank	Team	F1	P	R
1	TOKOFOU	**0.897**	0.907	0.896
2	dunder_mifflin	0.891	0.907	0.878
3	NARNIA	0.881	0.900	0.879
...	...	...	...	...
5	advex	0.858	0.882	0.864
...	...	...	...	...
8	HunterSpeechLab	0.736	0.874	0.684

Table 3.2: Top English: Evaluation

Rank	Team	F1	P	R
1	R00	**0.781**	0.842	0.763
*	iCompass	0.748	0.978	0.737
2	HunterSpeechLab	0.741	0.804	0.700
3	advex	0.728	0.809	0.753

Table 3.3: Top Arabic: Evaluation
* late submission

3.2 Dataset

There are three language datasets from the workshop: the English set, the Arabic set, and the Bulgarian set. In addition, we made a fourth dataset that combines all three

Rank	Team	F1	P	R
1	advex	**0.837**	0.860	0.861
2	HunterSpeechLab	0.817	0.819	0.837
3	majority_baseline	0.792	0.742	0.855

Table 3.4: Top Bulgarian: Evaluation

tweet _no	tweet_text	q1_ label	q2_ label	q3_ label	q4_ label	q5_ label	q6_ label	q7_ label
38	Corona Virus isn't real, the government just wants everyone inside for two weeks so they can change the batteries in the birds.	yes	yes	no	no	no	no	no

Table 3.5: Tweet example

languages. Each language dataset contains training, dev, test, and gold sets. Training and dev sets contain tweet no., text, q1 label, q2 label to q7 label. Test sets contain tweet no. and text, while gold contains the same tweet no. and text of the test set and the tweet question labels. Table 3.5 is an example of an English training dataset. Initially, the dataset came from (Alam et al., 2020). The workshop creators added additional tweets in Arabic and a new Bulgarian dataset.

Label	Detail
No	not harmful
No	joke or sarcasm
Yes	panic
Yes	xenophobic, racist, prejudices or hate-speech
Yes	bad cure
Yes	rumor or conspiracy
Yes	other

Table 3.6: Matrix

	English	Arabic	Bulgarian
Training	755	2534	2998
Dev	53	520	350
Test	418	1000	357

Table 3.7: Dataset summary: Number of tweets

Figure 3.1 is the English training dataset, which contains 755 tweets and is preprocessed (replaces links with "URL"), and tweet no. 385 is removed because the tweet was long enough to crash the training (not enough RAM). The English dev dataset contains 53 tweets and is preprocessed. English test dataset contains 418 with no preprocessing. Figure 3.1 shows unbalanced labels in all questions, the majority of question 1 is yes, while question 2's yes is the majority. Thus, most verifiable factual claims are not claims of false information (if there is no evidence, it is more likely to be false).

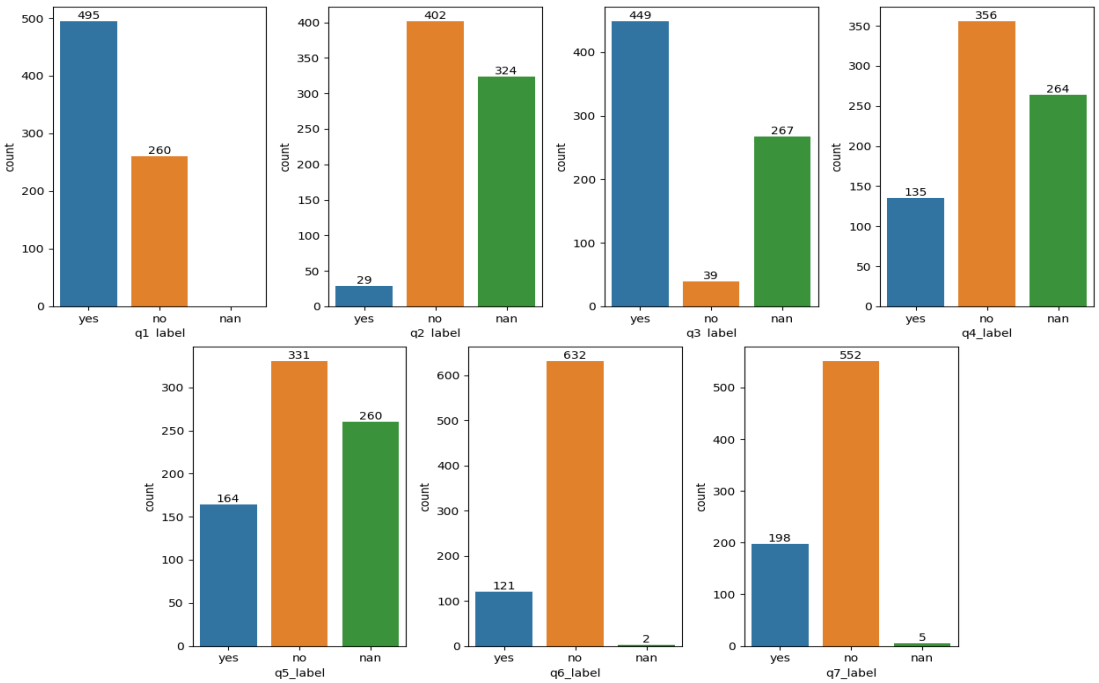


Figure 3.1: English training set

Figure 3.2 is the Arabic training dataset containing 2534 tweets (originally 198

tweets) and is preprocessed. The Arabic dev dataset contains 520 tweets and is preprocessed. The Arabic test dataset contains 1000 tweets. Figure 3.2 shows unbalanced labels in most questions. Most questions have the same pattern as the English dataset except for question 7. Thus, most tweets in Arabic require government attention.

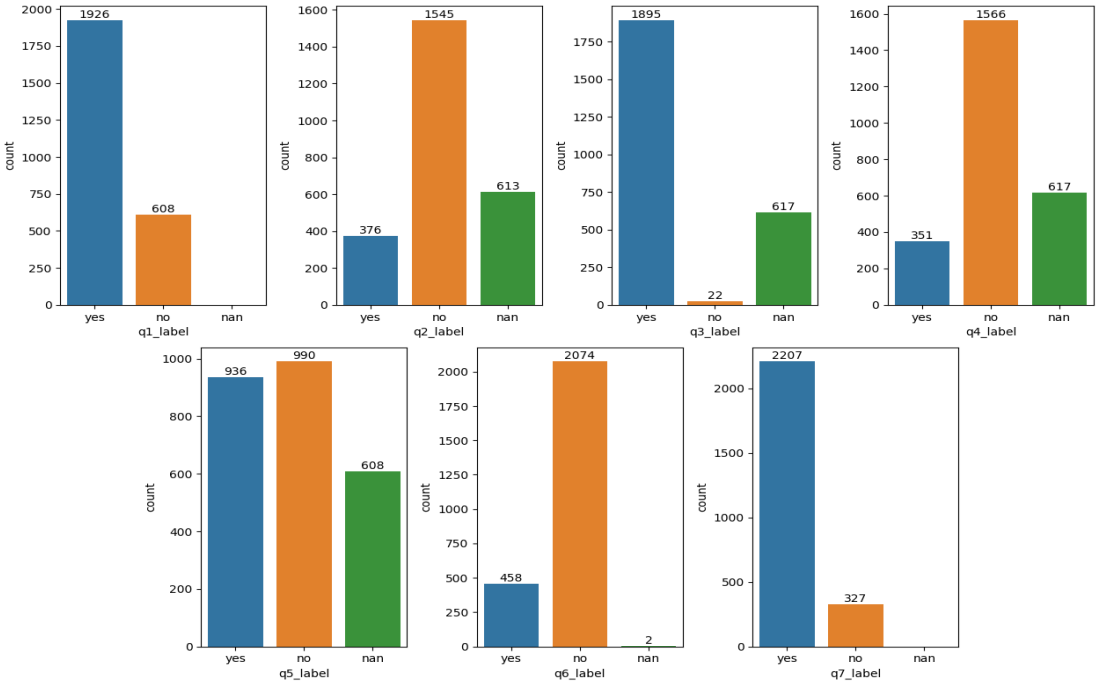


Figure 3.2: Arabic training set

Figure 3.3 is the Bulgarian training dataset containing 2998 tweets. Tweets 1769 and 2227 were removed because they were long enough to crash (not enough RAM). The dev dataset contains 350 tweets with preprocessing. The testing dataset contains 357 tweets. Figure 3.3 shows unbalanced labels in all questions, the same as the English dataset.

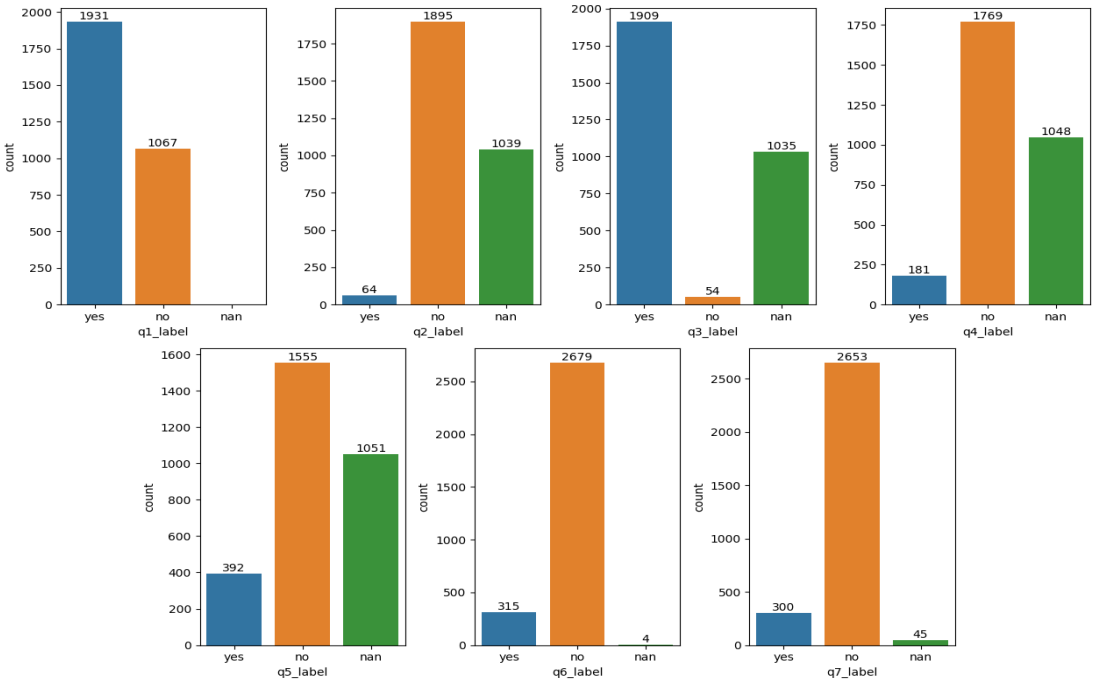


Figure 3.3: Bulgarian training set

3.3 Methodology

This section aims to use LLMs to finetune different tasks (in three languages).

The original TOKOFOU system uses six pre-trained BERT models in English to perform the task and uses majority voting aggregation. These six models are:

- Model 1: BERTWEET is based on RoBERTa, which is trained on 850M tweets and an additional 23M tweets (COVID-19 related) were trained (Nguyen, Vu & Nguyen, 2020).
- Model 2: CT-BERT uses BERT large model and is trained on 97M tweets related to COVID-19 from crowdbreak (collection tweets of a specific topic on Twitter) (Müller, Salathé & Kummervold, 2020).
- Model 3-6: TWEETEVAL is based on RoBERTa and is trained on 60M tweets. TOKOFOU uses four fine-tuned TWEETEVAL models to detect: hate speech,

emotion, irony and offensive. It is fine-tuned by adding a dense layer (connection of the previous layer) and training on specific classification tasks (Barbieri, Camacho-Collados, Neves & Espinosa-Anke, 2020).

TOKOFOU is trained for "15 epochs on batches of 16 tweets, using the AdamW optimizer with a learning rate of $3 \cdot 10^{-5}$ and a weight decay of 0.01" (Tziafas et al., 2021, p. 121). The TOKOFOU team proposed an ensemble model where the outcomes are determined by a majority vote (6 models). Pre-processing is done by replacing URL links (https:example.com) with "URL". Therefore, the URL links do not affect the weight of the model. Post-processing involves using the question 1 result to determine if questions 2–5 become nan (if q1 is no, then q2-q5 are nan) (Tziafas et al., 2021).

3.3.1 TOKOFOU Architecture

Figure 3.4 shows the training of TOKOFOU. Step 1 of the process is that each pre-trained transformer algorithm is trained/tested for 20 epochs (TOKOFOU_T) and fine-tuned for each question with the AdamW optimizer. The final algorithm models use the best epoch model (based on the F1 score). Figure 3.5 shows Step 2 of the process: the test dataset is inputted to each model and outputs a prediction. In step 3, the outcome is determined by the predicted outcomes of each model by majority rule. The majority-voted answer is then post-processed.

TOKOFOU_M

TOKOFOU_M is TOKOFOU with existing code (multilingual model) but did not submit their score into the competition. Table 3.8 shows TOKOFOU_M models used.

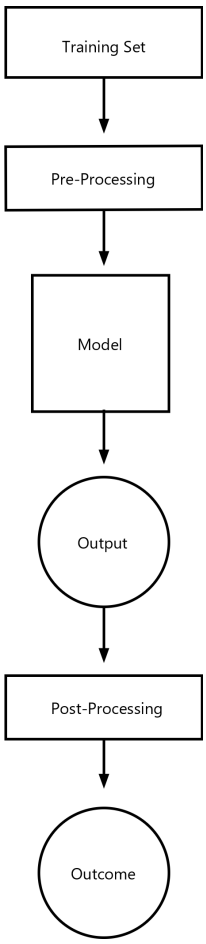


Figure 3.4: TOKOFOU training architecture

Training data is pre-processed, and a batch of 16 tweets each is trained and tested until the whole training set is looked through (epochs). The prediction is post-processed, and the loss is calculated. This sequence is completed 20 times (20 epochs).

TOKOFOU_T

TOKOFOU_T uses the same architecture as TOKOFOU, using 20 epochs instead of 15 and additional models.

- TOKOFOU_T All is finetuned for all three languages (trained on the combined dataset).
- TOKOFOU_T English is finetuned for English.
- TOKOFOU_T Arabic is finetuned for Arabic.

- TOKOFOU_T Bulgarian is finetuned for Bulgarian.

Table 3.8 shows an overview of models used in TOKOFOU_T.

Model Type - Language	Models
TOKOFOU - English	del-covid, vinai-covid, cardiffnlp-offensive, cardiffnlp-hate, cardiffnlp-emotion, cardiffnlp-irony
TOKOFOU_M - All	multi-bert, multi-xlm, multi-microsoft, multi-sentiment, multi-toxic
TOKOFOU_T - All	multi-bert, multi-xlm, multi-microsoft, multi-sentiment, multi-toxic, multi-xlm-roberta-base-snli-mnli-anli-xnli
TOKOFOU_T - English	del-covid, cardiffnlp-offensive, cardiffnlp-hate, cardiffnlp-emotion, cardiffnlp-irony, cardiffnlp-tweet, vinai-tweet
TOKOFOU_T - English (Multi)*	multi-xlm, multi-xlm-roberta-base-snli-mnli-anli-xnli, multi-verdict-classifier, multi-sentiment, multi-toxic
TOKOFOU_T - Arabic	arabic-xlm-r-base, arabert, multi-toxic, bert-base-arabic, multi-xlm-roberta-base-snli-mnli-anli-xnli, camelbert-mix-msa
TOKOFOU_T - Bulgarian	multi-bert, multi-xlm, multi-microsoft, multi-sentiment, multi-toxic, multi-verdict-classifier

Table 3.8: Overview model
Each language in TOKOFOU_T uses different models and is trained by their respective language, * TOKFOU_T English Multi is early model

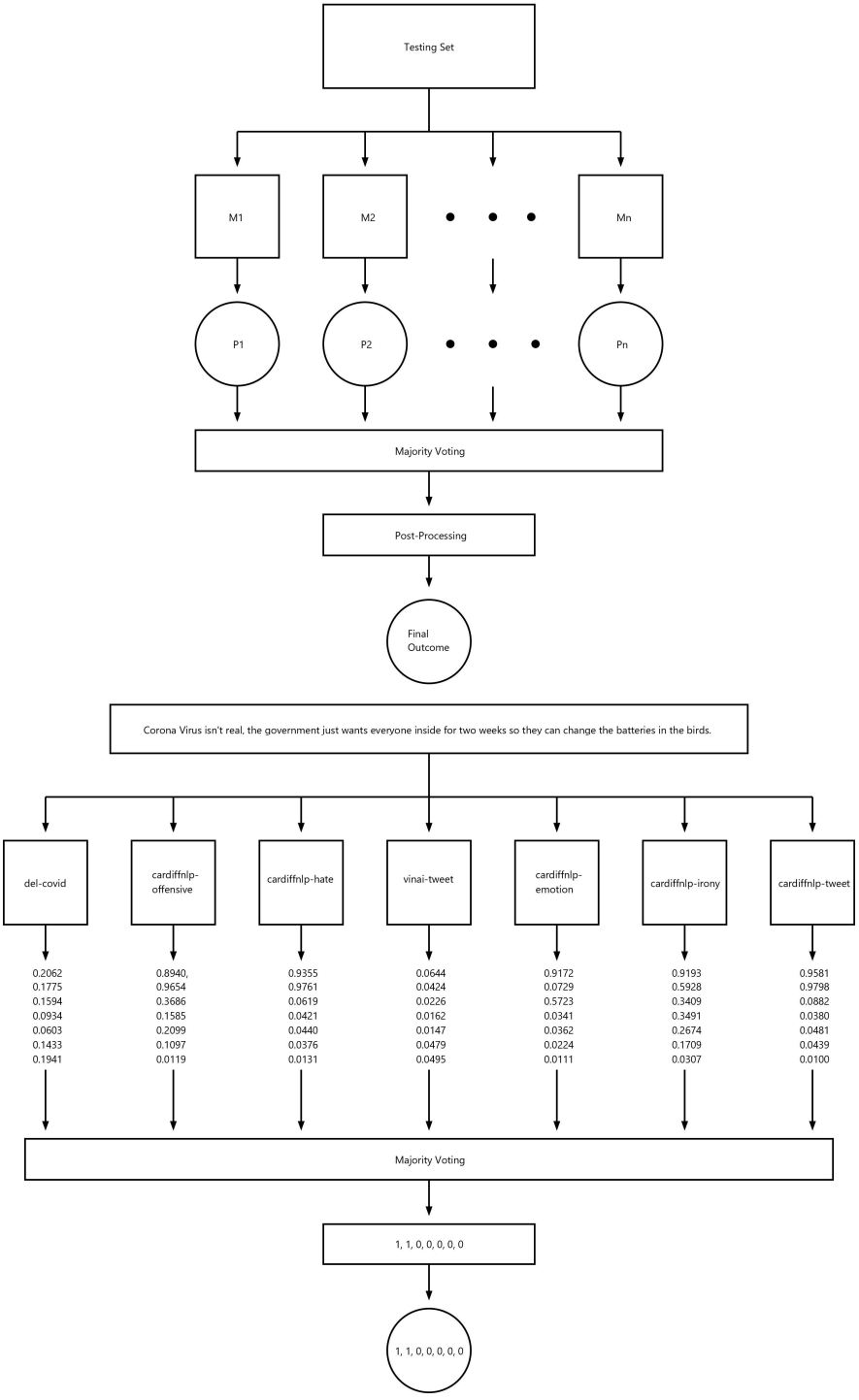


Figure 3.5: TOKOFOU testing architecture

Model (M) and Prediction (P). All testing set is sent to each model to predict the Q1-7. The outcome is determined by majority rules which are then post-processed to output the final labels. Example tweet: "Corona Virus isn't real, the government just wants everyone inside for two weeks so they can change the batteries in the birds." (Note: 1 is Yes and 0 is No).

3.3.2 BERT

Bidirectional Encoder Representations from Transformer (BERT) is a semi-supervised learning model researched by Google in 2018 (Devlin, Chang, Lee & Toutanova, 2018). BERT has two main steps: First, the models are trained on unlabeled data (pre-training) for different tasks. The next step is using the labelled data to tune the pre-trained parameters. Each task has different fine-tuned models. The pre-training involves two tasks: masked Language Modeling (masking 15% of random tokens from the input) and Next Sentence Prediction (to understand the relationship between sentences A and B by choosing 50% of sentence B to be a random sentence). Fine-tuning involves each task's input and output parameters being tuned end to end (converting). BERT was trained on BooksCorpus (800M words) and English Wikipedia (2,500M words - text only). Figure 3.6 shows how a general BERT works. The sentence is first tokenized (the [CLS] token is the beginning of the input, each word is a token, and new sentences start with the [SEP] token). The tokens are changed to ids and sent to BERT (multiple encoders) with segment embedding (distinguish different sentences) and position embedding (numbered order) (Devlin et al., 2018). For TOKOFOU, the output of BERT is sent to a Neural Network (fine-tuned NN), in which each question has its own parameters.

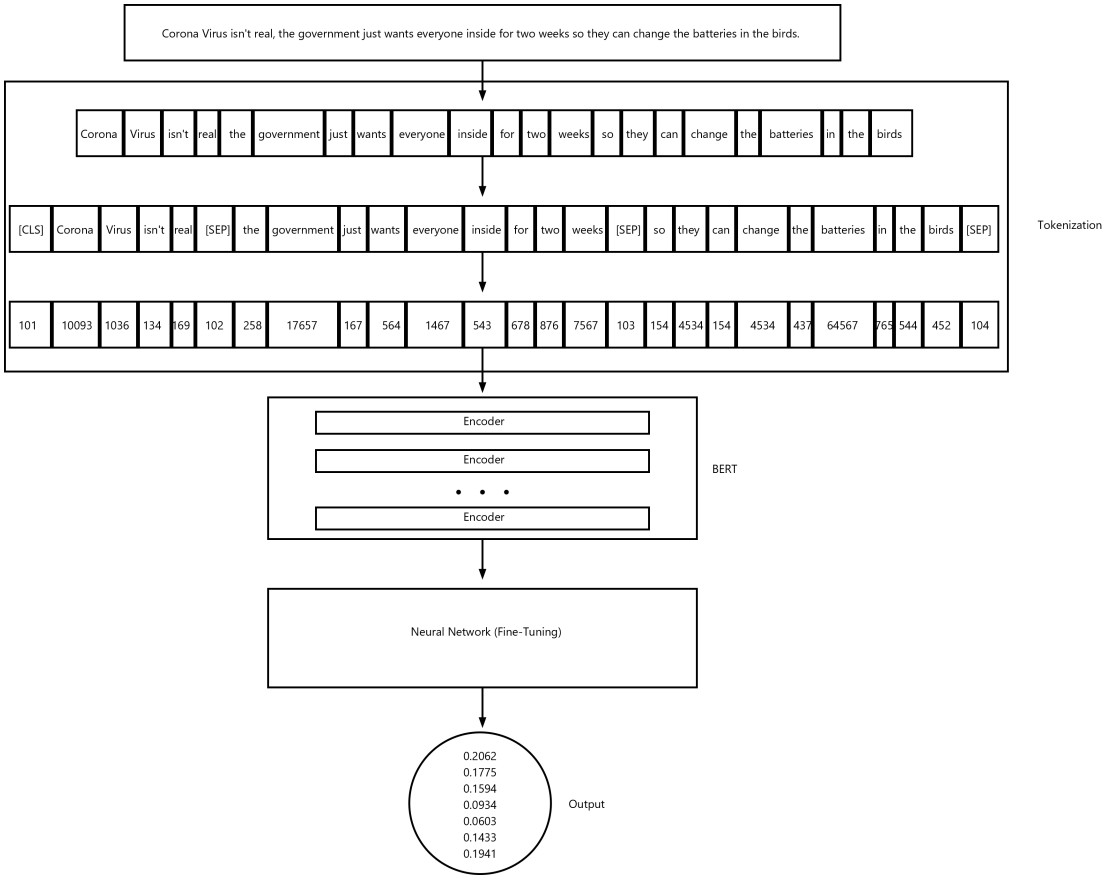


Figure 3.6: BERT example

This is an example of how a general BERT works. First, the sentence is tokenized into ids (random in this example) and sent to BERT (multiple encoders). The output of BERT is sent to a Neural Network (fine-tuned NN), in which each question has its parameters.

3.3.3 RoBERTa

RoBERTa is an improved version of BERT by training more data, bigger batches, and longer sequences, removing Next Sentence Prediction (NSP) and modifying the masking behaviour when training (Liu et al., 2019). Dynamic masking is used in the pre-training stage instead of a static mask. This is done by duplicating the training data 10 times, each sequence with 10 maskings over 40 epochs. There is a slight performance improvement by changing the model input format and replacing the NSP

with Full-Sentences (each input is a maximum of 512 tokens long). The text encoding is modified by training with a larger Byte-Pair-Encoding (BPE) size of 50K (without pre-processing) instead of 30K (with pre-processing), which adds 15M-20M parameters to the BERT base and BERT large (Liu et al., 2019).

3.3.4 Multilingual BERT

Multilingual BERT is used in the improved version of TOKOFOU: BERT base multilingual cased (multi-bert) is from a BERT that is pre-trained on 104 languages from Wikipedia (lowercased, undersampling large Wikipedia languages, and oversampling small Wikipedia languages) (*BERT multilingual base model (cased)*, n.d.). The BERT-based multilingual uncased sentiment (multi-sentiment) is based on BERT and RoBERTa. It is trained to identify trolling, aggression, and cyberbullying on Twitter and Youtube in English, Hindi, and Bengali (Mishra, Prasad & Mishra, 2020). XLM-RoBERTa (multi-xlm) is a multilingual version of RoBERTa and is pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages (Conneau et al., 2019). Multilingual MiniLm (multi-microsoft) is based on the BERT architecture, uses the XLM-R tokenizer, and is trained on XML-R (Deep self-attention Distillation) (Wang et al., 2020). Multilingual Toxic XLM-RoBERTa (multi-toxic) is an XLM-R based model that is trained to identify toxicity and pre-trained on Wikipedia comments (English, French, Spanish, Italian, Portuguese, Turkish, and Russian) (Hanu & Unitary team, 2020). XLM RoBERTa base snli mnli anli xnli is a XLM-R based model that is trained on the Stanford Natural Language Inference (SNLI) Corpus, the Multi-Genre Natural Language Inference (MNLI) corpus, the Adversarial Natural Language Inference (ANLI) and the Cross-Lingual NLI (XNLI) Corpus (*XLM RoBERTa Base Snli Mnli Anli Xnli*, n.d.). Multilingual Verdict Classifier is a RoBERTa model trained on 2,500 multilingual verdicts from Google Fact Check Tools, which are translated into 65 languages using

Google Translate (*Verdict Classifier*, n.d.).

3.3.5 Arabic BERT

In the improved TOKOFOU, several Single Language Transformers are used: BERT-based Arabic (bert-base-arabic), which is trained on Arabic Wikipedia (Safaya, Abdullatif & Yuret, 2020). Arabert is a BERT-based model that is trained on Oscar unshuffled and filtered, Arabic Wikipedia, the Arabic corpus (1.5B words), the Open Source International Arabic News (OSIAN) corpus, and Assafir news articles (Antoun, Baly & Hajj, 2020). CAMelBERT mix msa (Computational Approaches to Modelling Language Lab BERT) is a collection of BERT models that were trained on Arabic datasets (modern standard Arabic, dialectal Arabic, and classical Arabic) (Inoue, Alhafni, Baimukan, Bouamor & Habash, 2021). AdaSL (arabic-xlm-r-base) is an XLM-R-based model that is trained on a 5 million Arabic tweet corpus (1 million per dialect) (Mekki, Mahdaouy, Berrada & Khoumsi, 2022).

3.3.6 GPT-3

In this section, we tried using the testing data on Chatgpt 3.5 and fined tuned GPT-3 Ada using the training used in TOKOFOU_T. GPT-3 is a generative pre-training from openAI and uses the same architecture as GPT-2. GPT-3 has 175 billion parameters and is trained on 499 billion tokens from filtered CommonCrawl, Webtext, Book1, Book2 and Wikipedia (Brown et al., 2020)

ChatGPT3.5

chatGPT3.5 is a fined-tune version (supervised and reinforcement learning from human trainers) of GPT3 from openAI.

By using this following prompt: Here is a claim, "TWEET", can you answer the following 7 questions in a very concise manner:

- Q1: is this Verifiable Factual Claim? - Answer Yes if there is claim which is supported by evidence such as statistics, verifiable, factual information, examples or personal testimony. No if there is no evidence.
- Q2: does this claim contain False Information? - Yes if there is claim of false information.
- Q3: Is this claim of the Interest to General Public? - Yes if the claim relates to potential cures, updates on number of cases, on measures taken by governments or dis- cussing rumors and spreading conspiracy theories.
- Q4: is this claim Harmful? Yes if the claim is harmful to the society / person(s) / company(s) / product(s).
- Q5: is this claim in Need of Verification? Yes if a fact checker is needed verify the claim.
- Q6: is this claim harmful to society? No if not harmful; No if a joke or sarcasm; Yes if it can create panic; Yes if it has xenophobic, racist, prejudices or hate-speech; Yes if bad cure; Yes if rumor or conspiracy; Yes if others.
- Q7: does this claim Require attention? Yes if the claim needs to be looked by government department.

and with the following settings:

1. model = gpt-3.5-turbo
2. temperature = 0
3. message = [’role’: ’user’, ’content’ : PROMPT]

Finetune GPT3 Ada

GPT3 Ada is one of four models that can be fine-tuned. Ada is suited for classification problems. Although it scores slightly lower than the other three, it is the fastest and cheapest model. The fine-tune of GPT3 Ada is using the existing training used in TOKOFOU and modifying the training data to fit openAI requirements: {"prompt:":"TWEET ->", "completion":" no\nnan\nnan\nnan\nnan\nyes\nno\n ###"} ("->" means prompt ending separator and "###" means completion ending separator) additionally using openAI's default parameters. The cost of training Ada is $0.0004 USD per 1K tokens, and using it costs $0.0016 USD per 1K tokens. This is cheaper than chatGPT3.5 turbo ($0.002 USD per 1K tokens) and does not need to add questions in each tweet, as seen in the chatGPT3.5 prompt.

3.4 Evaluation

All training and testing were done on a laptop with an Intel i7-8750h (6 cores, 12 threads) at 3.00 GHz and 16 GB of RAM. A few rows in some datasets were removed due to limited RAM (explained in section 3.2). Table 3.8 is an overview of what models we used in the TOKOFOU_T. In the analysis of the original TOKOFOU code, there was experimental code for other languages: multi-bert, multi-xlm, multi-xnli (large model), multi-microsoft, multi-sentiment and multi-toxic. In the TOKOFOU_T, additional models were added: multi-xlm-roberta-base-snli-mnli-anli-xnli, multi-verdict-classifier, arabic-xlm-r-base, arabert, bert-base-arabic and camelbert-mix-msa.

3.4.1 Results

Below are the key TOKOFOU types:

- TOKOFOU - Original TOKOFOU score.

Model	Name	Training	Parameters	Language/s
GPT-3 turbo	Generative Pre-trained Transformer	499 Billion tokens (CommonCrawl, WebText2, Books1, Books2, Wikipedia) plus chat conversations (chatgpt3.5 turbo)	175B	95
del-covid (CT-BERT)	COVID-Twitter-BERT	97M tweets (1.2B words)	340M (BERT-Large base)	English
vinai-covid/vinai-tweet	BERTWEET	English tweets and Covid tweets, 850 tweets (16B words)	110M (BERT-base)	English
cardiffnlp-...	TweetEval	60M tweets	355M (RobERTa-base)	English
multi-bert	BERT multilingual base model (cased)	Wikipedia	110M (BERT-base)	104
multi-xlm	XLM-RoBERTa (cross-lingual language model)	2.5TB filtered common crawl data	110M (BERT-base)	100
multi-microsoft	MiniLM	NAN	21M	16
multi-sentiment	NAN	24.1K tweets and youtube	110M (BERT-base)	NAN
multi-toxic	Detoxify	Wikipedia talk page edits and CivilComments (organisation)	550M (XLM-RoBERTa-base)	7
multi-xlm-roberta-base-snli-mnli-anli-xnli	NAN	SNLI, MNLI, ANLI and XNLI	270M (XLM-RoBERTa-base)	13
multi-verdict-classifier	Multilingual Verdict Classifier	2.5K verdicts from Google Fact Check Tools API and translated	270M (XLM-RoBERTa-base)	65
arabic-xlm-r-base	AdaSL	5M tweets	270M (XLM-RoBERTa-base)	Arabic
arabert	AraBERT	8.6B words (OSCAR unshuffled and filtered, Arabic Wikipedia, 1.5B Arabic Corpus, OSIAN Corpus and Assafir news articles	136M	Arabic
bert-base-arabic	ArabicBERT	8.2B words (OSCAR, Arabic Wikipedia, Other)	110M (BERT-base)	Arabic
multi-camelbert-mix-msa	CAMeLBERT Modern Standard Arabic	17.3B words and PATB dataset	NAN	Arabic

Table 3.9: Model summary
NAN: does not specify or not found

- TOKOFOU_M - Original TOKOFOU multilingual code that was not submitted.
- TOKOFOU_T - Improved TOKOFOU.

English

Table 3.10 shows the original TOKOFOU score as the top performing system by 0.6% of the following best (dunder_mifflin) and 1.6% (NARNIA) for second best. The top-scoring questions were 4, 5, and 7 from TOKOFOU. Question 6 has a difference of 3.2% between TOKOFOU and dunder_mifflin, which shows TOKOFOU's weakness in determining whether the tweet harms society. Question 2 has a difference of 2.2% between advex, majority_baseline and TOKOFOU, which shows some weakness in labelling if there is a claim of false information in tweets. Question 3 has a 1.4% difference between majority_baseline and TOKOFOU, which shows some weakness in determining if the tweet contains interesting content for the public. In question 1, TOKOFOU_T English performed 0.5% better than TOKOFOU. Overall TOKOFOU_T English ranked fourth. TOKOFOU_All ranked fifth, and in question 3, it performed 2.6% better than TOKOFOU_T English. TOFOKOU_M performed 10th out of 12 systems. Therefore, it was possibly not submitted because of the low score.

Rank	Team	Overall			Q1	Q2	Q3	Q4	Q5	Q6	Q7
		F1	P	R	F1	F1	F1	F1	F1	F1	F1
1	TOKOFOU	0.897	0.907	**0.896**	0.835	0.913	0.978	**0.873**	**0.882**	0.908	**0.889**
2	dunder_ mifflin	0.891	0.907	0.878	0.807	0.923	0.966	0.868	0.852	**0.940**	0.884
3	NARNIA	0.881	0.900	0.879	0.831	0.925	0.976	0.822	0.854	0.909	0.849
4	TOKOFOU _T English	0.880	0.893	0.887	**0.840**	0.915	0.961	0.834	0.853	0.886	0.873
**	ChatGPT 3.5 turbo	0.878	0.889	0.882	0.659	***0.934***	0.980	***0.873***	***0.913***	0.915	0.869
**	GPT3 Ada English	0.872	0.890	0.880	0.843	0.900	0.975	0.805	0.833	0.900	0.848
5	TOKOFOU _T All	0.871	0.894	0.887	0.798	0.912	0.982	0.821	0.851	0.884	0.848
**	GPT3 Ada All	0.864	0.885	0.859	***0.852***	0.883	0.955	0.807	0.813	0.894	0.847
6	InfoMiner	0.864	0.897	0.848	0.819	0.886	0.946	0.841	0.803	0.884	0.867
7	advex	0.858	0.882	0.864	0.784	**0.927**	0.987	0.858	0.703	0.878	0.866
8	Lang-Research-LabNC	0.856	**0.909**	0.827	0.842	0.873	0.914	0.829	0.792	0.894	0.849
9	TOKFOU _T English (multi)	0.855	0.884	0.869	0.833	0.904	0.961	0.773	0.811	0.879	0.824
10	TOKOFOU _M	0.854	0.878	0.870	0.819	0.910	0.950	0.774	0.827	0.878	0.820
*	majority_ baseline	0.830	0.786	0.883	0.612	**0.927**	**1.000**	0.770	0.807	0.873	0.821
*	ngram_ baseline	0.828	0.819	0.868	0.647	0.904	0.992	0.761	0.800	0.873	0.821
11	Hunter-Speech-Lab	0.736	0.874	0.684	0.738	0.822	0.824	0.744	0.426	0.878	0.720
12	spotlight	0.729	0.907	0.676	0.813	0.822	0.217	0.764	0.701	0.905	0.877
*	random_ baseline	0.496	0.797	0.389	0.552	0.480	0.457	0.473	0.423	0.563	0.526

Table 3.10: English: Evaluation
* baseline model, ** GPT model (later evaluation)

Arabic

Table 3.11 shows the TOKOFOU_T Arabic has the best score by 0.2% of R00 (single language), 3.5% of the best multilingual system other than TOKOFOU (iCompass) and 2.6% of TOKOFOU_M. Advex, who came sixth, has the best question score than TOKOFOU_T (1.4% in Q2, 3.7% in Q3 and 3.0% in Q4) and would have performed better overall if Q7 had a higher score. TOKOFOU_T Arabic question 5 score performed best by 5% of R00 and 7.3% of advex. Question 5 and 7 performance in all systems is low (Q5 best is 64.6% and 69.0% in Q7). TOKOFOU_T All, ranked third and in question, performed better than TOKOFOU_T Arabic by 6.2%.

Rank	Team	Overall			Q1	Q2	Q3	Q4	Q5	Q6	Q7
		F1	P	R	F1	F1	F1	F1	F1	F1	F1
1	TOKOFOU _T Arabic	**0.783**	0.833	**0.779**	0.725	0.787	0.944	0.829	**0.646**	0.888	0.659
2	R00	0.781	**0.842**	0.763	0.843	0.762	0.890	0.799	0.596	**0.912**	0.663
3	TOKOFOU _T All	0.776	0.824	0.770	0.787	0.766	0.926	0.812	0.628	0.882	0.630
4	TOKOFOU _M	0.757	0.813	0.763	0.757	0.756	0.925	0.782	0.556	0.870	0.651
5	iCompass	0.748	0.784	0.737	0.797	0.746	0.881	0.796	0.544	0.885	0.585
6	Hunter-Speech-Lab	0.741	0.804	0.700	0.797	0.729	0.878	0.731	0.500	0.861	**0.690**
7	advex	0.728	0.809	0.753	0.788	**0.821**	**0.981**	**0.859**	0.573	0.866	0.205
8	InfoMiner	0.707	0.837	0.639	**0.852**	0.704	0.774	0.743	0.593	0.698	0.588
**	ngram_ baseline	0.697	0.741	0.716	0.410	0.762	0.950	0.767	0.553	0.856	0.579
**	ChatGPT 3.5 turbo	0.682	0.800	0.696	0.730	0.742	0.926	0.799	0.523	0.886	0.171
**	GPT3 Ada All	0.671	0.793	0.643	0.785	0.702	0.865	0.722	0.516	0.837	0.270
9	Damascus-Team	0.664	0.783	0.677	0.169	0.754	0.915	0.783	0.583	0.857	0.589
**	majority_ baseline	0.663	0.608	0.751	0.152	0.786	0.981	0.814	0.475	0.857	0.579
10	spotlight	0.661	0.805	0.632	0.843	0.703	0.792	0.647	0.194	0.828	0.620
**	GPT3 Ada Arabic	0.608	0.768	0.574	0.732	0.617	0.774	0.634	0.435	0.850	0.212
**	random_ baseline	0.496	0.719	0.412	0.510	0.444	0.487	0.442	0.476	0.584	0.533

Table 3.11: Arabic: Evaluation
* baseline model ,** GPT model (later, evaluation)

Bulgarian

Table 3.12 shows TOKOFOU_T Bulgarian is the second best performer by 1.5% by advex. TOKOFOU_T Bulgarian and TOKOFOU_M are best at classifying question 1 by 1.2% compared to HunterSpeechLab and 6.2% compared to advex. Question 2 TOKOFOU_T Bulgarian came fourth out of 9 models or third out of 6 (teams) by 1.7% compared to advex. TOKOFOU_T Bulgarian's question 3 came third or tied with HunterSpeechLab and performed 1.8% worse than majority_baseline and 1.2% worse than advex. Question 4, TOKOFOU_T Bulgarian came third/tie with majority_baseline and performed worse by 1.3% compared to HunterSpeechLab. TOKOFOU_T Bulgarian performed second in question 5 by 1.5% compared to advex. TOKOFOU_T All ranked 5th and performed slightly worse than TOKOFOU_M by 0.4% and TOKOFOU_T Bulgarian by 1.6%. TOKOFOU_T All performed better than TOKOFOU_T Bulgarian in question 6 by 1.2%.

Rank	Team	Overall			Q1	Q2	Q3	Q4	Q5	Q6	Q7
		F1	P	R	F1	F1	F1	F1	F1	F1	F1
1	advex	**0.837**	**0.860**	**0.861**	0.887	**0.955**	0.980	0.834	**0.819**	**0.678**	**0.706**
2	TOKOFOU_T Bulgarian	0.823	0.857	0.854	**0.949**	0.938	0.968	0.822	0.804	0.615	0.664
3	Hunter-Speech-Lab	0.817	0.819	0.837	0.937	0.943	0.968	**0.835**	0.748	0.605	0.686
4	TOKOFOU_M	0.811	0.829	0.850	**0.949**	0.938	0.968	0.815	0.761	0.606	0.643
5	TOKOFOU_T All	0.807	0.844	0.821	0.921	0.914	0.944	0.820	0.766	0.627	0.657
*	majority_baseline	0.792	0.742	0.855	0.876	0.951	**0.986**	0.822	0.672	0.606	0.630
*	ngram_baseline	0.778	0.790	0.808	0.909	0.919	0.949	0.803	0.631	0.606	0.630
**	ChatGPT 3.5 turbo	0.739	0.817	0.701	0.898	0.761	0.895	0.653	0.655	0.667	0.642
**	GPT3 Ada Bulgarian	0.718	0.814	0.688	0.799	0.809	0.824	0.707	0.634	0.605	0.646
6	spotlight	0.686	0.844	0.648	0.832	0.926	0.336	0.669	0.687	0.650	0.700
**	GPT3 Ada All	0.677	0.818	0.627	0.729	0.742	0.754	0.658	0.610	0.605	0.640
7	InfoMiner	0.578	0.826	0.505	0.786	0.749	0.419	0.599	0.556	0.303	0.631
*	random_baseline	0.496	0.768	0.400	0.594	0.502	0.470	0.480	0.399	0.498	0.528

Table 3.12: Bulgarian: Evaluation
* baseline model, ** GPT model (later evaluation)

All languages

Table 3.13 shows when TOKOFOU_T All is trained on All languages (combining English, Arabic, and Bulgarian) and tested with the combined testing dataset. The results show high scores in most questions and the F1 score. Questions 5 and 7 seem to

Overall			Q1	Q2	Q3	Q4	Q5	Q6	Q7
F1	P	R	F1	F1	F1	F1	F1	F1	F1
0.823	0.850	0.819	0.812	0.863	0.946	0.812	0.732	0.834	0.759

Table 3.13: All language: Evaluation

be on the low side, as seen in Figure 3.11.

Swapping languages

Model Type - Learned Dataset Language					
	TOKOFOU_T All - Trained All languages*	TOKOFOU - Trained on English	TOKOFOU_T English (multi) - Trained on English	TOKOFOU_T Arabic - Trained on Arabic	TOKOFOU_T Bulgarian - Trained on Bulgarian
Test language	F1 Score	F1 Score	F1 Score	F1 Score	F1 Score
English	**0.871**	***0.884***	***0.855***	0.726	**0.847**
Arabic	0.776		0.642	***0.783***	0.657
Bulgarian	0.807		0.683	0.703	*0.823*
All	*0.823*		0.713	0.755	0.744

Table 3.14: Swap language: Evaluation
Bold is best model F1 score, Italic is learned language, * combined with three lanaguages

In this section, the experiment compares if testing a model in a different language dataset from its trained language dataset gives similar performance. In Table 3.14, English performs best in most models except for TOKOFOU_T Arabic. When Arabic is trained, the best F1 score is 78.3% compared to 64.2% in TOKOFU_T English, 65.7% in TOKOFOU_T Bulgarian and 77.6% in TOKOFOU_T All. When Bulgarian trained, the best F1 score is 84.7% compared to 68.3% in TOKOFOU_T English, 70.3% in TOKOFOU_T Arabic, 82.3% in TOKOFOU_T Bulgarian and 80.7% in TOKOFOU_T All. When All languages are trained, the best F1 score is 82.3% compared to 71.3% in TOKOFOU_T English, 75% in TOKOFOU_T Arabic and 74.4% in TOKOFOU_T Bulgarian.

TOKOFOU_T Bulgarian produced their best score in the English dataset even if the model was not fine-tuned further in English. This shows that the multi-language model performs best in English because it has been trained in English more. The fine-tuning of the English and Bulgarian datasets does change the performance, as seen in 3.14. Fine-tuning English is 0.8% better than without fine-tuning. When fine-tuned in Bulgarian, it is 14.9% better than without fine-tuned. In addition, when fine-tuned for English or Bulgarian, they perform similarly (1.5% difference) when tested in Arabic. This could be because of the different models used. TOKOFOU_T All has the second best performance in different languages, by 1.3% worse in TOKOFOU English, 1.6% better in TOKOFOU_T English, 0.7% worse in Arabic, and 1.6% worse in Bulgarian.

TOKOFOU_T All tested on the All language set to have a similar F1 score to TOKOFOU_T Bulgarian of 0.823 in All and 0.823 in Bulgarian test, which is a coincidence as question F1 is different as shown in Table 3.15

Question	All F1 score	Bulgarian F1 score
Q1	0.812	0.949
Q2	0.863	0.938
Q3	0.946	0.968
Q4	0.812	0.822
Q5	0.732	0.804
Q6	0.834	0.615
Q7	0.759	0.664

Table 3.15: All and Bulgarian F1 comparison

TOKOFOU_T All has the most robust F1 score, which came second in all three languages.

3.4.2 GPT3 results

GPT3.5 turbo

In some tweets chatGPT 3.5 sometimes does not answer with a yes or no, but "potentially ..." or "cannot determine without further information" or "yes, if...". These answers would be modified according to Table 3.16

GPT answer	Modified answer	Modified Q1, Q6 and Q7 answer
not specified	nan	no
potentially...	yes	yes
not indicated	nan	no
not necessarily	no	no
as an ai language...	nan	no
no information provided	nan	no
no,...	no	no
yes,...	yes	yes

Table 3.16: Result modification

Table 3.10 shows the result of the English test set with GPT models. ChatGPT3.5 is ranked fifth and has better or identical individual scores as the best algorithm in questions 2, 4, and 5, but poor performance in question 1. Perhaps the understanding of question 1 differs from human to LLMs, or the question is poorly written. Thus low overall score. ChatGPT's performance was surprising to show a high overall F1 score in English without any training from the dataset. When changing question 1 for chatGPT to understand better, it performed lower. The chatGPT Arabic score showed lower overall performance than other algorithms. Although most question F1 scores were high on average, question 7's score is the worst out of all the algorithms (significantly low overall score). Therefore, there could be a translation misunderstanding when using

English questions. The chatGPT Bulgarian score showed low overall performance. Questions 1 and 2 showed good performance, but others showed low performance.

Finetuned GPT3 Ada

The results from finetuned GPT3 Ada in Table 3.10 showed lower performance than chatGPT by 0.6%, but a significantly better F1 score in question 1 by 18.4%. Question 1 in GPT3 Ada All scored 0.8% better than fine-tuned English and is the top score out of the other algorithms in English. The finetuned Arabic in Table 3.11 score showed the worst performance of all algorithms (0.74% below chatGPT), especially in questions 5 and 7. Questions 1 and 7 showed some improvement over chatGPT. The Bulgarian in Table 3.12 shows the finetune Bulgarian. It ranked lower than chatGPT (by 0.21%) but performed better in questions 2, 4, and 7. The finetuned version of GPT3 Ada did not improve the overall score. However, in some cases, it improved individual questions such as question 1 in English, question 7 in Arabic, and question 4 in Bulgarian. GPT3 Ada All performed slightly worse than fine-tuned English by 0.8% and fine-tuned Bulgarian by 4.1%. Surprisingly GPT3 Ada All performed better than fine-tuned Arabic by 6.3%.

3.5 Discussion and Conclusion

Overall, TOKOFOU_T All is ranked one on average across all three languages (out of 8 multilingual methods in Table 3.17). TOKOFOU's weakness is in questions 2, 3, and 6, both seen in non-modified and modified models. This could be because of the type of model used or trained on.

Rank	Team	English F1	Arabic F1	Bulgarian F1	Average
1	TOKOFOU_T All	0.871	0.776	0.807	0.818
2	TOKOFOU_M	0.854	0.757	0.811	0.807
3	advex	0.858	0.728	0.837	0.806
4	ChatGPT3.5 Turbo	0.878	0.682	0.739	0.766
5	HunterSpeechLab	0.736	0.741	0.817	0.765
6	GPT3 Finetuned Ada All	0.864	0.671	0.677	0.737
7	InfoMiner	0.864	0.707	0.578	0.716
8	spotlight	0.729	0.661	0.686	0.692

Table 3.17: Multilingual methods: Top average F1 score

Eight algorithms were tested in all three languages. Table 3.17 shows TOKOFOU_T All and the seven other algorithms with F1 score and rank in each language. TOKOFOU_T All ranked nearly top in almost all languages except for Bulgarian, and it is the best multilingual algorithm (average score). Advex ranked top for Bulgarian but fourth for English and Arabic, TOKOFOU_M ranked fourth in English, second in Arabic and third in Bulgarian. HunterSpeechLab ranked around third in almost all languages except for English (sixth), InfoMiner ranked low in Arabic and Bulgarian except for English (third), and spotlight ranked low in all languages.

In general, Q5 and 7 have low performance in most languages. This could be the label's lack of clarity or the dataset it was fine-tuned on needed to be more balanced, as seen in Section 3.2.

TOKOFOU_T All shows high scores in all languages. TOKOFOU ranked first in English but did not participate in Arabic and Bulgarian. This could be because of their low scores when using the dev dataset, as there were planned multilingual models. The TOKOFOU_T (3.8 overview of models) used five extra models (multi-xlm-roberta.., arabert, arabic-xlm-r-base, bert-base-arabic and camelbert-mix-msa) that were fine-tuned on the given dataset. Of the three languages, Bulgarian performance

was lower than the advex by 1.4% but 4.5% better in Arabic. Compared with other models, TOKOFOU_T All scores have no weak performances in all three languages shown in Table 3.17

ChatGPT's performance in the English test set was surprising because of the high score (fifth rank and one rank below TOKOFOU_T English) without any training. Question 1's score could have performed better. Therefore, changing how the question is written may improve the score, but after several attempts, there was no improvement (lower score).

The fine-tuning of GPT3 Ada improved question 1 significantly but gave a slightly lower score in other questions. To try to improve the score, the training hyperparameters were modified (such as using the "babbge" model, changing the n_epoches and batch size), and in addition, the training prompt was modified (including questions similar to the chatGPT prompt). The changes did not improve the scores, resulting in a lower overall score.

Chapter 4

Explainability on False Information Detection

This chapter provides a study of explainability on the outputs of TOKOFOU and GPT3 system. Explainable AI (XAI) is a way for humans to understand/explain the results of an AI system. By having XAI, it would bring forwards unexpected or undesirable output, increase trust from users, follow the General Data Protection Regulation (GDPR), improve the system over iterations, discover manipulation efficiently, and know why the system chose that result (Adadi & Berrada, 2018; Confalonieri, Coba, Wagner & Besold, 2020; Longo, Goebel, Lecue, Kieseberg & Holzinger, 2020; Meske, Bunde, Schneider & Gersch, 2020; Das & Rad, 2020; Samek & Müller, 2019). Thus, XIA systems allow it to "explain what it has done, what it is doing now, and what will happen next" (Gunning et al., 2019, p. 4)

Two main types of methods to explainability are: designing an algorithm built to be already interpretable and creating a system to explain the unexplained model without modifying the original model, called post-hoc explanation (Adadi & Berrada, 2018; Confalonieri et al., 2020). The AI methods, such as decision trees, fuzzy logic, and naive bayes, belong to the first type as they are naturally explainable without additional

system modification or performance degradation (Longo et al., 2020; Das & Rad, 2020; Angelov, Soares, Jiang, Arnold & Atkinson, 2021). As AI models get more complex and especially by using neural networks, it is harder to interpret/explain them and do risk assessment/analysis (Adadi & Berrada, 2018; Longo et al., 2020). We'll now review two well-known post-hoc explanation methods.

The first method is Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016), which explains the prediction from a model by finding the relationships to words in sentences or patches of images. LIME shows the relationship by calculating the weights for the prediction (positive) or against the prediction (negative), which allows the user to understand the prediction from the model and whether to trust it. One example of a model that predicts right, but the reason for the prediction is untrustworthy, is shown in the example of Figure 4.1, which classifies whether the document is Christian or Atheism. Several words ("Posting", "Host", and "Re") in the document made no connection to Atheism because it was found that 99% of documents that have "Posting" are Atheism.

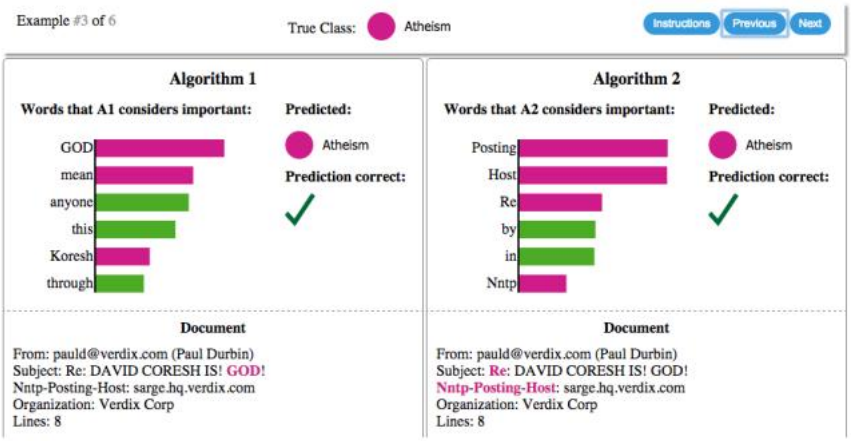


Figure 4.1: Untrustworthy example
From *"Why Should I Trust You?" Explaining the Predictions of Any Classifier* (Ribeiro et al., 2016, p. 2)

Additionally, LIME can explain images by using super-pixels (groupings of pixels) to give weight to each class. This will give insights into what group of pixels influences

the model's outcome. An example of a model explaining whether the image is of a husky or wolf, LIME showed that the model classified huskies when snow is in the background and wolf when there is no snow (Ribeiro et al., 2016).

The second method is SHapely Additive exPlanations (SHAP), Lundberg and Lee (2017) proposed SHAP which measures the importance of features in predicting black-box systems. SHAP values are a measurement of each feature. This calculates the value from the base value to an output ("They explain how to get from the base value that would be predicted if we did not know any features to the current output" (p. 5)). There are several types of approximation methods: Linear SHAP (linear model), Low-Order SHAP, MAX SHAP, Kernal SHAP (LIME + SHAP values), and Deep SHAP (DeepLIFT + SHAP values).

4.1 Explainablity in TOKOFOU

In this section, we use LIME and SHAP to explain four tweets in the test dataset for TOKOFOU. We only do this for English because we are unfamiliar with the other two languages.

4.1.1 LIME

A tweet is inputted to LIME, LIME masks the tweet, and it is sent to TOKOFOU, and the final probability of each question is sent back to LIME (to calculate) to explain TOKOFOU's prediction.

Here we list four examples and discuss their interpretations in LIME.

Tweet ID: 927 The original text: "The media and Republicans are trying to give Trump credit for the coronavirus vaccine, but the truth is the vaccines were in development months before Operation Warp Speed. https://t.co/6qOnyCGf2H via @politicususa"

Q1	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**
months	but	months	Trump	the	Republicans	Republicans
were	6qOnyGf2H	the	Republicans	Trump	Trump	Trump
the	Trump	Trump	months	months	vaccines	vaccines
Trump	Republicans	were	are	coronavirus	media	coronavirus
Republicans	vaccines	Republicans	the	Republicans	development	months
Y 0.97, N 0.03	Y 0.02, *N 0.98*	*Y 0.87*, N 0.13	*Y 0.62*, N 0.38	Y 0.37, *N 0.63*	Y 0.18, *N 0.82*	Y 0.33, *N 0.67*

Table 4.1: Tweet ID: 927 Top 5 keywords LIME
Italic is ground truth, Cyan is negative keywords

Table 4.1 shows[1] an overview of the predictions of 7 questions and the top 5 keywords associated with the prediction according to LIME. In this case, TOKOFOU predicted it with 100% accuracy.

Question 1's probability of yes is 97% because of several positive keywords: 'months' 0.15, 'were' 0.14, 'Trump' 0.13 and 'Republicans' 0.12. The keywords show that if 'months', 'were', 'Trump' and 'Republicans' were removed (0.97 - 0.15 - 0.14 - 0.13 - 0.12 = 0.43). TOKOFOU would have predicted a question 1 yes probability of 43% vs 97%. It is seen similarly in questions 3, 4, 5 and 6. In most questions, if 'Trump' and 'Republican' were in the tweet, it would be a more highly predicted Yes to the questions. 'vaccines' and 'development' affect its probability for questions 6 and 7 no's. However, the rest has 'the' (stop word) as a negative, which does not give an insightful explanation. Question 2 has no explanation, possibly because there is a probability of only 0.2% of being yes. More keywords are found in the Appendix A

[1]Figure B.1 in page 156 has the original output from LIME.

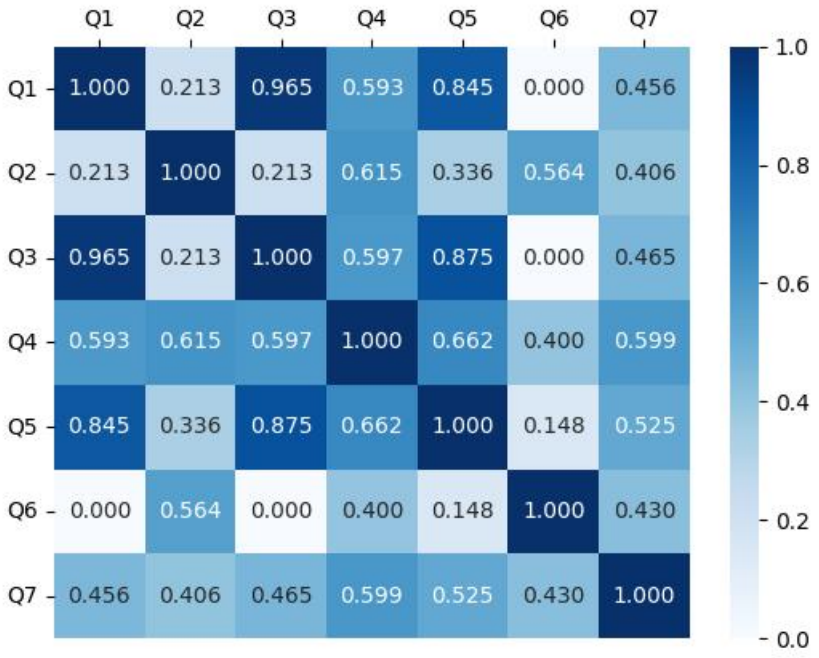


Figure 4.2: Tweet ID: 927 Spearman's rank correlation coefficient

Figure 4.2 shows Spearman's rank correlation coefficient of questions in tweet 927. This compares the top 20 keywords in each question with each other and shows any similarities between the questions. The highest coefficient is in questions 1 and 3 at 0.965, the second highest in questions 3 and 5 at 0.875, and the third is in questions 1 and 5 at 0.845. Questions 1, 3, and 5 have high similarities in their top 20 keywords. The questions may relate to each other. Question 4 has some relation (0.5 above) to most questions. Thus, there are some key relationships. Question 6 has the lowest average coefficient of most questions.

Tweet ID: 1020 The original text: "Why vaccine nationalism can't work for most people, in one graphic. The stuff needed to make, distribute & administer vaccines comes from around the world. Trade barriers make scaling up harder. From paper by @SorescuSilvia, @jlopezgonzalez1 & Andrenelli https://t.co/61HwWOlmy7

https://t.co/c3szOeDyiq"

Q1	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**
Trade	t	Trade	can	can	can	can
work	Trade	vaccines	Trade	work	Why	Why
vaccine	barriers	vaccine	barriers	Trade	make	Trade
barriers	scaling	work	work	t	around	work
vaccines	vaccine	barriers	t	barriers	nationalism	t
Y 0.79, *N 0.21*	Y 0.01, *N 0.99*	Y 0.61, *N 0.39*	Y 0.09, *N 0.91*	Y 0.05, *N 0.95*	Y 0.02, *N 0.98*	Y 0.10, *N 0.90*

Table 4.2: Tweet ID: 1020 Top 5 keywords LIME
Italic is ground truth, Cyan is negative keywords

Table 4.2 shows[2] an overview of the predictions of 7 questions and the top 5 keywords associated with the prediction according to LIME. In this case, TOKOFOU predicted it with 71% accuracy (assume nan as no).

Question 1 probability of yes is 79%, and question 3 is 61%. Both questions' keywords are: 'Trade', 'vaccines', 'vaccine' and 'work'. Questions 4, 5, 6 and 7 have 'can' and 'Why' as top keywords. Question 2 does not show enough information to give meaningful word insights because of the low Yes probability. Due to question 1 being yes, question 3 did not change to nan, thus resulting in 71% accuracy and not 100% accuracy.

Figure 4.3 shows Spearman's rank correlation coefficient of questions in tweet 1020. This compares the top 20 keywords in each question with each other and shows any similarities between the questions. The highest coefficient is questions 1 and 3 at 0.929, the second highest is questions 4 and 7 at 0.595, and the third is questions 5 and 7 at 0.561. Questions 1, 3, and 4 have high similarities among their top 20 keywords, as seen in tweet 927. In addition, the output of questions 1 and 3 is misclassified. Questions 4, 5, and 7 have some relation (0.5 above) to each other, so there is some relation to the

[2]Figure B.2 in page 157 has the original output from LIME.

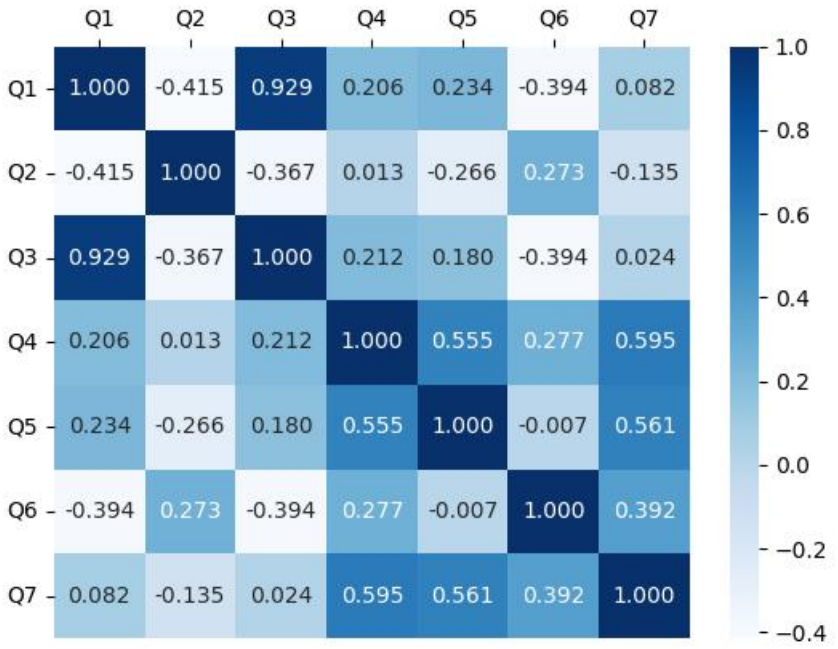


Figure 4.3: Tweet ID: 1020 Spearman's rank correlation coefficient

question. Questions 2 and 6 have the lowest average coefficients of most questions.

Tweet ID: 1066 The original text: "The media are doing everything they can to erase any and all contributions the Trump administration did to create and distribute the COVID-19 vaccines in record time. https://t.co/gxDPjmDBZB"

Table 4.3 shows[3] an overview of the predictions of 7 questions and the top 5 keywords associated with the prediction according to LIME. In this case, TOKOFOU predicted it with 57% accuracy (assume nan as no).

The most common keywords are 'record', 'erase', 'Trump' and 'media'. The keywords are seen in all questions except for question 2 (low yes probability). Questions 5, 6 and 7's output is wrong, which could mean the weights of the keywords were lower even though they have the exact common positive keywords.

[3]Figure B.3 in page 158 has the original output from LIME.

Q1	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**
record	erase	record	erase	erase	erase	erase
erase	media	erase	Trump	record	media	Trump
Trump	record	Trump	media	Trump	Trump	to
doing	Trump	doing	record	doing	distribute	administration
are	COVID	the	are	are	vaccines	did
Y 0.89, N 0.11	Y 0.04, *N 0.96*	*Y 0.61*, N 0.39	*Y 0.70*, N 0.30	*Y 0.23*, N 0.77	*Y 0.40*, N 0.60	*Y 0.34*, N 0.66

Table 4.3: Tweet ID: 1066 Top 5 keywords LIME
Italic is ground truth, Cyan is negative keywords

Figure 4.4: Tweet ID: 1066 Spearman's rank correlation coefficient

Figure 4.4 shows Spearman's rank correlation coefficient of questions in tweet 1066. This compares the top 20 keywords in each question with each other and shows any similarities between the questions. The highest coefficient is questions 1 and 3 at 0.898, the second highest is questions 3 and 5 at 0.889, and the third is questions 1 and 5 at

0.844. Questions 1, 3, and 5 are highly similar to their top 20 keywords, as seen in tweet 927. In addition, the output of questions 5, 6, and 7 is misclassified. Questions 4, 5, and 7 have some relation (0.5 above) to each other and are seen in tweet 1020. Questions 2 and 6 have the lowest average coefficient of most questions, as seen in tweet 1020.

Tweet ID: 962 The original text: "The US has a significant interest in sharing the outcomes of successful Covid vaccine research to protect both global health and our own health security. Here's how the Biden administration can expand global manufacturing of vaccines. https://t.co/HtmB0KwY00"

Q1	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**
US	US	US	US	US	expand	vaccines
has	expand	interest	sharing	vaccines	vaccines	successful
expand	manufacturing	expand	successful	has	US	expand
manufacturing	administration	has	vaccines	sharing	manufacturing	sharing
health	interest	health	outcomes	interest	successful	vaccine
Y 0.83, *N 0.17*	Y 0.01, *N 0.99*	Y 0.77, *N 0.23*	Y 0.01, *N 0.99*	Y 0.01 , *N 0.99*	Y 0.01, *N 0.99*	Y 0.03, *N 0.97*

Table 4.4: Tweet ID: 962 Top 5 keywords LIME
Italic is ground truth, Cyan is negative keywords

Table 4.4 shows[4] an overview of the predictions of 7 questions and the top 5 keywords associated with the prediction according to LIME. In this case, TOKOFOU predicted it with 71% accuracy (assume Nan is no).

Since most questions have a low yes probability, there is insufficient information to give insights into questions 2, 4, 5, 6 and 7. Questions 1 and 3 have similar keywords but different weights: ' US', 'has', 'expand' and 'health'. Questions 1 and 3 differ because 'interest' (negative) is a higher rank in question 1. If question 1 were right, it would have 100% accuracy. 'manufacturing' and 'expand' are positive in question 1

[4]Figure B.4 in page 159 has the original output from LIME.

but negative in other questions. Thus the model may misclassify these keywords (also seen in question 3).

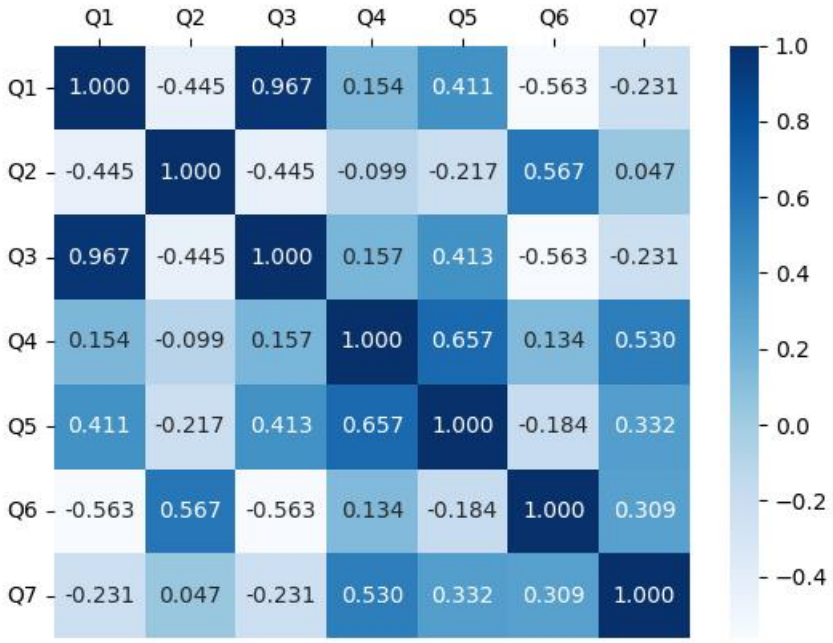


Figure 4.5: Tweet ID: 962 Spearman's rank correlation coefficient

Figure 4.5 shows Spearman's rank correlation coefficient of questions in tweet 962. This compares the top 20 keywords in each question with each other and shows any similarities between the questions. The highest coefficient is questions 1 and 3 at 0.967, the second highest is questions 4 and 5 at 0.657, and the third is questions 2 and 6 at 0.567. Questions 1, 3, and 4 have high similarities among their top 20 keywords, as seen in tweet 1020. Additionally, the output of questions 1 and 3 is misclassified. Questions 2, 4, 5, and 6 have some relation (0.5 above) to each other. Questions 2 and 6 have the lowest average coefficient of most questions, also seen in tweets 1020 and 1066. Questions (1,6), (3,6), (1,2), and (2,3) have a high negative coefficient, meaning that the ranking of keywords is the opposite.

4.1.2 SHAP

A sentence is inputted to SHAP, and SHAP then outputs masked words in sentences to TOKOFOU. TOKOFOU's prediction is sent back to SHAP to create an explanation of TOKOFOU.

Tweet ID: 927 The original text: "The media and Republicans are trying to give Trump credit for the coronavirus vaccine, but the truth is the vaccines were in development months before Operation Warp Speed. https://t.co/6qOnyCGf2H via @politicususa"

Q1	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**
months before	co	Republicans	Republicans	Republicans	Republicans	Republicans
Republicans	and	vaccines	are	are	are	are
vaccines	Republicans	months before	Trump	Trump	were	Trump
were	but	are	co	co	vaccines	co
are	via	were	media	months	Trump	give
Y 1, N 0	Y 0, *N 1*	*Y 1*, N 0	*Y 1*, N 0	Y 0, *N 1*	Y 0, *N 1*	Y 0, *N 01*

Table 4.5: Tweet ID: 927 Top 5 keywords SHAP
Italic is ground truth, Cyan is negative keywords

Table 4.5 shows[5] an overview of the predictions of 7 questions and the top 5 keywords associated with the prediction according to SHAP.

Questions 4, 5, 6, and 7 have identical top keywords: 'Republican', 'are' and 'Trump'. Question 1 and 3 has identical keywords ('months before', 'Republicans', 'vaccines', 'are' and 'were') and similar weights to them. Question 2 negative keywords are stop words such as 'and', 'but', 'via' and 'co' (part of a link).

[5]Figure B.7 in page 162 has the original output from SHAP.

Tweet ID: 1020 The original text: "Why vaccine nationalism can't work for most people, in one graphic. The stuff needed to make, distribute & administer vaccines comes from around the world. Trade barriers make scaling up harder. From paper by @SorescuSilvia, @jlopezgonzalez1 & Andrenelli https://t.co/61HwWOlmy7 https://t.co/c3szOeDyiq"

Q1	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**
vaccines	.	vaccines	can	can	can	can
from	barriers	from	’	’	’	’
comes	Trade	administer	t	t	vaccines	t
Trade barriers	graphic	comes	work	work	world	world
administer	scaling	.	vaccine nationalism	Silvia	administer	work for
Y 1, *N 0*	Y 0, *N 1*	Y 1, *N 0*	Y 0, *N 1*	Y 0, *N 1*	Y 0, *N 1*	Y 0, *N 1*

Table 4.6: Tweet ID: 1020 Top 5 keywords SHAP
Italic is ground truth, Cyan is negative keywords

Table 4.6 shows[6] an overview of the predictions of 7 questions and the top 5 keywords associated with the prediction according to SHAP.

Question 1 and 3 has similar top positive keywords: ’ vaccines’, ’from’, ’comes’, ’administer’, and the top similar negative is ’graphic’. Question 2 is primarily negative: the top keywords are ’full stop’, ’barriers’, ’Trade’ and ’graphic’. Questions 4,5,7 similar positive keywords are ’can/’t’ and ’work’, while question 6 does have ’work’ as a keyword. Question 5 does not have ’vaccine’ and ’nationalism’ to affect the prediction. Seeing question 6 keywords, we see that some of questions 1 and 3 (’vaccines’ and ’administer’) are positive, while question 6 is negative. These keywords may be misclassified (positive should be negative).

[6]Figure B.8 in page 163 has the original output from SHAP.

Tweet ID: 1066 The original text: "The media are doing everything they can to erase any and all contributions the Trump administration did to create and distribute the COVID-19 vaccines in record time. https://t.co/gxDPjmDBZB"

Q1	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**
record	erase	record	erase	erase	erase	erase
.	media	erase	media	media	media	to
erase	co	.	are	to	are	media
time	t	in	Trump	t	t	to
administer	https	vaccines	to	co	*in*	https
Y 1, N 0	Y 0, *N 1*	*Y 1*, N 0	*Y 1*, N 0	*Y 0*, N 1	*Y 0*, N 1	*Y 0*, N 1

Table 4.7: Tweet ID: 1066 Top 5 keywords SHAP
Italic is ground truth, Cyan is negative keywords

Table 4.7 shows[7] an overview of the predictions of 7 questions and the top 5 keywords associated with the prediction according to SHAP.

Most questions' top keywords are 'erase', 'media', and 'COVID'. However, question 1 and 4 has '-19', 'vaccines', 'record' and 'administration did' as top keywords. The tokenisation in SHAP splits COVID-19 into two words which can impact the final verdict of keyword weight. Question 5, 6 and 7 is wrong, and this is because of low keyword weights.

Tweet ID: 962 The original text: "The US has a significant interest in sharing the outcomes of successful Covid vaccine research to protect both global health and our own health security. Here's how the Biden administration can expand global manufacturing of vaccines. https://t.co/HtmB0KwY00"

Table 4.8 shows[8] an overview of the predictions of 7 questions and the top 5 keywords associated with the prediction according to SHAP.

[7]Figure B.9 in page 164 has the original output from SHAP.
[8]Figure B.10 in page 165 has the original output from SHAP.

Q1	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**
US	.	US	sharing	US	expand	US
expand	US	expand	US	sharing	and	vaccine
manufacturing	expand	manufacturing	successful	has	global	successful
has	and	has	has	The	US	sharing
sharing	has	how	vaccine	successful	vaccine	expand
Y 1, *N 0*	Y 0, *N 1*	Y 1, *N 0*	Y 0, *N 1*	Y 0 , *N 1*	Y 0, *N 1*	Y 0, *N 1*

Table 4.8: Tweet ID: 962 Top 5 keywords SHAP
Italic is ground truth, Cyan is negative keywords

Most questions have 'US' as a positive except for questions 2 and 6. Questions 1 and 3 have the exact top 4 keywords and rankings. Therefore, they may have some correlations with each other. Questions 1, 4 and 5 have 'sharing' as a top negative keyword and significantly affect the outcome. Similar to question 1 in LIME, 'expand' could have been misclassified, resulting in a low negative weight to the overall outcome.

4.1.3 Comparison

The comparison of LIME and SHAP on Tweet 927 (100% accuracy) and 962 (71% accuracy).

Tweet ID: 927 Comparison

Question 1	
LIME	**SHAP**
months	months before
were	Republicans
the	vaccines
Trump	were
Republicans	are

Question 2	
LIME	**SHAP**
but	co
6qOnyGf2H	and
Trump	Republicans
Republicans	but
vaccines	via

Question 3	
LIME	**SHAP**
months	Republicans
the	vaccines
Trump	months before
were	are
Republicans	were

Question 4	
LIME	**SHAP**
Trump	Republicans
Republicans	are
months	Trump
are	co
the	media

Question 5	
LIME	**SHAP**
the	Republicans
Trump	are
months	Trump
coronavirus	co
Republicans	months

Question 6	
LIME	**SHAP**
Republicans	Republicans
Trump	are
vaccines	were
media	vaccines
development	Trump

Question 7	
LIME	**SHAP**
Republicans	Republicans
Trump	are
vaccines	Trump
coronavirus	co
months	give

Table 4.9: Tweet 927: Comparison LIME and SHAP
Cyan is negative keywords

Table 4.9 shows a comparison of LIME[9] and SHAP[10] keywords of tweet 927.

In question 1, three keywords: 'months' (SHAP includes 'before'), 'Republican' and 'were' are both found in LIME and SHAP, although two in different ranks. 'months' seems to affect question 1, the most and seen in LIME and SHAP at the identical rank.

[9]Figure B.1 in page 156 has the original output from LIME.
[10]Figure B.7 in page 162 has the original output from SHAP.

One negative keyword found from LIME to have an effect in question 1 is 'the', which is ranked higher in LIME than SHAP ('is the').

In question 2, two keywords: 'but' and 'Republicans', are both found in LIME and SHAP, although in different ranks. 'but' is the top rank in LIME and affects negatively, but SHAP has 'co', which is part of a link. Additionally, part of the same link is seen in rank 2 in LIME. LIME ranks 'Trump' as the top 3, but SHAP ranked it low. SHAP seems to have letters in the link to be the most influential negative and stop words such as 'and', 'but' and 'via'.

In question 3, three keywords: 'months' ('months before' in SHAP), 'Republicans' and 'were' are both found in LIME and SHAP, although in different ranks. 'the' is seen in LIME as second rank, while ranked low in SHAP. 'vaccines' is ranked 11th while 3rd in SHAP.

In question 4, three keywords: 'Trump', 'Republicans' and 'are' are found in LIME and SHAP, although in similar ranks. LIME ranks 'Trump' higher than 'Republican' in LIME, while the opposite in SHAP.

In question 5, two keywords: 'Trump' and 'Republicans', are both found in LIME and SHAP. However, in different ranks, especially 'Republicans' is ranked much lower in LIME verse SHAP. 'the' is a negative word ranked highest in LIME while 'co' is ranked 4th in SHAP. 'coronavirus' is ranked 4th while SHAP does not have it ranked.

In question 6, three keywords: 'Republicans', 'Trump', and 'vaccines' are both found in LIME and SHAP, both LIME and SHAP ranked 'Republicans' top influential but ranked 'Trump' and 'vaccines' higher in LIME verse SHAP. 'media' and 'development' is ranked 6th and 7th in SHAP.

In question 7, two keywords: 'Republicans' and 'Trump', are found in both LIME and SHAP, 'Republicans' is both ranked top and 'Trump' is ranked one rank different.

From these comparisons, SHAP tends to use the letter ('co') from the link to a higher negative rank, putting more stop words higher than LIME. LIME's top negative

words are 'the' while 'co' for SHAP. Both 'Trump' and 'Republicans' are seen in the top 5 more in LIME than SHAP, and SHAP tends only to have 'Republicans' in the top 5 ranks.

Tweet ID: 962 Comparison

Question 1	
LIME	**SHAP**
US	US
has	expand
expand	manufacturing
manufacturing	has
health	sharing

Question 2	
LIME	**SHAP**
US	.
expand	US
manufacturing	expand
administration	and
interest	has

Question 3	
LIME	**SHAP**
US	US
interest	expand
expand	manufacturing
has	has
health	how

Question 4	
LIME	**SHAP**
US	sharing
sharing	US
successful	successful
vaccines	has
outcomes	vaccine

Question 5	
LIME	**SHAP**
US	US
vaccines	sharing
has	has
sharing	The
interest	successful

Question 6	
LIME	**SHAP**
expand	expand
vaccines	and
US	global
manufacturing	*US*
successful	vaccine

Question 7	
LIME	**SHAP**
vaccines	US
successful	vaccine
expand	successful
sharing	sharing
vaccine	expand

Table 4.10: Tweet 962: Comparison LIME and SHAP
Cyan is negative keyword

Table 4.10 shows a comparison of LIME[11] and SHAP[12] keywords of tweet 962.

[11]Figure B.4 in page 159 has the original output from LIME.
[12]Figure B.10 in page 165 has the original output from SHAP.

Questions 1 and 3 outputs are misclassified. We can see that 'manufacturing' and 'expand' is seen negatively in other questions. These keywords might affect the outcome if the weights of the keywords were significant.

In question 1, four keywords: 'US', 'has', 'expand' and 'manufacturing' is found in both LIME and SHAP with similar/same ranking. 'sharing' is ranked higher in SHAP (5th) than LIME (7th) although not significantly. Overall, it is a better insight to question 1.

In question 2, there is not enough information in LIME to be reliable, but it has similar top keywords: 'US' and 'expand', which affect the most in question 1. SHAP shows that a full stop is the most influential to No, while LIME does not include it.

In question 3, there are three keywords: 'US', 'expand' and 'has' in both LIME and SHAP. 'US' and 'has' are in the same ranking, while 'expand' is one rank below in LIME. 'interest' is ranked significantly higher in LIME than SHAP, while 'how' is the opposite. 'manufacturing' is ranked lower in LIME (7th) than SHAP (3rd), which is the opposite for 'interest'. 'health' is not included in SHAP and is the 5th influential keyword.

In question 4, there are four keywords in both LIME and SHAP: 'US', 'sharing', 'successful' and 'vaccine. There are similarly ranked. Therefore, it is highly certain that these keywords are influential to Yes in question 4. 'has' on SHAP is ranked significantly higher than LIME (19th), and the opposite appears on 'outcomes'.

In question 5, there are three keywords: 'US', 'has' and 'sharing' both in LIME and SHAP. Those words are similarly ranked or ranked the same. LIME does not show insightful weights due to low yes probability. SHAP shows how influential each keyword is, although it is agreeable that 'US', 'sharing' and 'has' influence the most.

In question 6, three keywords: 'expand', 'vaccines' and 'US' are ranked top 5 in both LIME and SHAP. Words like 'global' and 'and' are ranked much lower or bottom in LIME than SHAP.

In question 7, four top 5 ranked words are found in LIME and SHAPE: ' vaccine', 'successful', 'expand', and 'sharing'. These words have a similar ranking. 'US' is ranked top in SHAP while 6th in LIME and is the top positive word.

Overall, there are more agreeable keywords seen in both LIME and SHAP. SHAP tends to rank stop words higher than LIME but did not see links or part of links in LIME and SHAP this time. 'US' tends to be the most influential keyword in LIME and SHAP in most questions.

4.2 Explainability of Finetune GPT3 Ada

The comparison of LIME and SHAP TOKOFOU and LIME GPT3 Ada

4.2.1 TOKOFOU and GPT3 Ada Comparison

Tweet ID: 927 Comparison

Question 1		
LIME	**LIME Ada**	**SHAP**
months	months	months before
were	development	Republicans
the	were	vaccines
Trump	in	were
Republicans	coronavirus	are

Question 2		
LIME	**LIME Ada**	**SHAP**
but	months	co
6qOnyGf2H	https	and
Trump	give	Republicans
Republicans	co	but
vaccines	development	via

Question 3		
LIME	**LIME Ada**	**SHAP**
months	months	Republicans
the	development	vaccines
Trump	were	months before
were	in	are
Republicans	coronavirus	were

Question 4		
LIME	**LIME Ada**	**SHAP**
Trump	vaccines	Republicans
Republicans	Republicans	are
months	months	Trump
are	trying	co
the	were	media

Question 5		
LIME	**LIME Ada**	**SHAP**
the	months	Republicans
Trump	Republicans	are
months	vaccines	Trump
coronavirus	were	co
Republicans	coronavirus	months

Question 6		
LIME	**LIME Ada**	**SHAP**
Republicans	vaccines	Republicans
Trump	Republicans	are
vaccines	truth	were
media	coronavirus	vaccines
development	trying	Trump

Question 7		
LIME	**LIME Ada**	**SHAP**
Republicans	vaccines	Republicans
Trump	Republicans	are
vaccines	truth	Trump
coronavirus	coronavirus	co
months	and	give

Table 4.11: Tweet 927 Comparison LIME, SHAP and LIME Ada
Cyan is negative keywords

Table 4.11 shows the comparison of TOKOFOU (LIME and SHAP) and the finetune GPT3 Ada (LIME). TOKOFOU had a 100% accuracy, and Ada had 86% accuracy (question 4 is misclassified).

Question 1 has two common keywords: 'months' and 'were'. Both Ada and

TOKOFOU show 'month' as the top keyword. Thus, it is the most likely word to help determine the outcome. 'were' is ranked in the middle, lower than TOKOFOU LIME and higher than SHAP.

Question 2, there is one common keyword in TOKOFOU SHAP and Ada LIME which is 'co', which is part of a link including 'https' (Ada) and '6q0nyGf2H' (TOKOFOU LIME).

In question 3, two common words are the same as in question 1, and Ada (LIME) is ranked the same as in question 1.

Question 4, Ada output is misclassified but has two common keywords: 'Republicans' and 'months', seen in Ada LIME and TOKOFOU LIME and ranked the same.

Question 5 has three common keywords: 'months', 'Republicans' and 'coronavirus'. 'months' is ranked higher than both TOKOFOU LIME and SHAP, but 'Republicans' is ranked higher in LIME and lower in SHAP.

Question 6, there are two common keywords which are 'vaccines' and 'Republicans'. 'vaccines' is ranked higher than TOKOFOU, but 'Republicans' is one rank lower than TOKOFOU.

Question 7 has three common keywords: 'vaccines', 'Republicans' and 'coronavirus'. Similar to question 6, 'vaccines' is ranked higher in Ada LIME, while 'Republicans' is ranked lower than TOKOFOU. 'coronavirus' is ranked the same in TOKOFOU LIME and Ada LIME.

Tweet ID: 962 Comparison

Question 1		
LIME	**LIME Ada**	**SHAP**
US	expand	US
has	manufacturing	expand
expand	Here	manufacturing
manufacturing	can	has
health	The	sharing

Question 2		
LIME	**LIME Ada**	**SHAP**
US	expand	.
expand	co	US
manufacturing	vaccines	expand
administration	https	and
interest	manufacturing	has

Question 3		
LIME	**LIME Ada**	**SHAP**
US	expand	US
interest	vaccine	expand
expand	US	manufacturing
has	manufacturing	has
health	The	how

Question 4		
LIME	**LIME Ada**	**SHAP**
US	vaccines	sharing
sharing	expand	US
successful	vaccine	successful
vaccines	US	has
outcomes	administration	vaccine

Question 5		
LIME	**LIME Ada**	**SHAP**
US	vaccines	US
vaccines	vaccine	sharing
has	expand	has
sharing	US	The
interest	sharing	successful

Question 6		
LIME	**LIME Ada**	**SHAP**
expand	vaccines	expand
vaccines	vaccine	and
US	HtmB0KwY00	global
manufacturing	sharing	US
successful	https	vaccine

Question 7		
LIME	**LIME Ada**	**SHAP**
vaccines	vaccines	US
successful	vaccine	vaccine
expand	US	successful
sharing	expand	sharing
vaccine	HtmB0KwY00	expand

Table 4.12: Tweet 962 Comparison LIME, SHAP and LIME Ada
Cyan is negative keyword

Table 4.12 shows the comparison of TOKOFOU (LIME and SHAP) and the finetune GPT3 Ada (LIME). Both algorithms had the same outcome and are 71% accurate (question 1 and 3 is misclassified).

Question 1 LIME Ada has two keywords in the top 5 of TOKOFOU: ' expand' and manufacturing'. Both are ranked higher than TOKOFOU (LIME and SHAP) and in the

same order. Thus, the keywords will most likely determine the outcome of question 1.

Question 2 Ada has two top 5 keywords opposite to TOKOFOU: 'expand' and 'manufacturing', although it answers correctly. The other three keywords ('co', 'vaccines' and 'https')are outside the top 5 TOKOFOU LIME and SHAP lists.

Question 3 Ada has three common keywords with TOKOFOU SHAP and two for LIME. 'expand' is ranked higher than TOKOFOU LIME and SHAP but lower for 'US' and 'manufacturing'.

Question 4 Ada has three common keywords 'US', 'vaccine' and 'vaccines'. 'vaccines' is higher ranked than TOKOFOU LIME and lower ranked for 'US'. In addition, the TOKOFOU SHAP keyword 'vaccine' is negative, while Ada LIME is positive.

Question 5 Ada has three common keywords 'vaccines', 'US' and 'sharing'. 'vaccines' is the top keyword, while 'US' is lower in Ada LIME than in TOKOFOU LIME and SHAP.

Question 6 Ada has two common keywords (one in LIME and one in SHAP), 'vaccines' and 'vaccine'. Ada LIME seems to have a link as the top keyword to determine the outcome rather than other questions and does not show in the top 5 in TOKOFOU.

Question 7 Ada has four common keywords 'vaccines', 'vaccine', 'US', and 'expand'. 'vaccines' is top in TOKOFOU LIME and Ada LIME, also 'vaccine' has the rank in both TOKOFOU SHAP and Ada LIME. Question 7 has the most common top 5 keywords in Ada and TOKOFOU.

4.3 Discussion and Conclusion

LIME and SHAP have shown insights into what keywords are associated with each question. LIME and SHAP often have the exact keywords, although at varying rankings, from a few to several rankings above or below. Most of the top positive and negative

keywords are the same in most questions. Questions 1 and 3 tend to have the exact keywords and similar rankings to each other, and the Spearman rank correlation coefficient shows this. Thus, there must be some relationship between the question and the answer. Additionally, questions 4, 5, and 7 have similar keywords and rankings but are less frequent than questions 1 and 3. However, question 7 differs from questions 4 and 5, although they are similar to the question (harmfulness and harmful to society).

Stop words are ranked higher in SHAP than LIME because the tokenization of the sentence may have various effects on the results. Additionally, tokenization in LIME and SHAP has different outcomes, such as the links in tweets that can be multiple words seen in Table 4.1, 4.5 and 4.7 or words are combined shown in 4.5.

The examples show no clear correlations between stop words and question results (Yes or No), as both can be negative or positive.

For future work, removing stop words in the tweet may have better results because stop words are frequently seen in LIME and SHAP's top 5 keywords. Using LIME tokenization on SHAP may also result in fairer outcomes and the exact keywords.

Chapter 5

Algorithms and Effects of Social Media

This chapter provides a review of the algorithms used by social media companies and their effects on society. Firstly, we explore the algorithms from four main social media services, including Facebook, Instagram, Twitter and Tiktok and their key features. Such algorithms can increase the spread of false information, and a better understanding of their inner working would be useful in mitigating the negative effects on society. Secondly, we review four main topics related to social effects: misinformation (Vosoughi et al., 2018; Fernández, Bellogín & Cantador, 2021), censorship (West, 2018; Patty, 2019), bias (such as echo chamber effects (Cinelli et al., 2021) and political biases (Chen et al., 2021)), and addiction (D'Arienzo, Boursier & Griffiths, 2019; Nazire Burcin Hamutoglu & Gezgin, 2020).

5.1 Algorithm Comparison of Social Media Companies

In this section, we compare the algorithms from four main social media services including Facebook, Instagram, Twitter and Tiktok and their key features. We first look at individual services and then provide a comparison in terms of their key features (see Table 5.1).

5.1.1 Facebook News Feed Algorithm

Facebook's news feed[1] is a tool that shows personalised stories from their social network (friends, family, interests, and current affairs). News feeds allow the user to see "meaningful and informative stories". Therefore, the user gets what they want and will likely react to the stories, thus staying on the platform longer. Adjustments can be made to the feed to suit users' preferences by adjusting favourites, follows, and snoozing.

As of 2022, Facebook has the following main features:

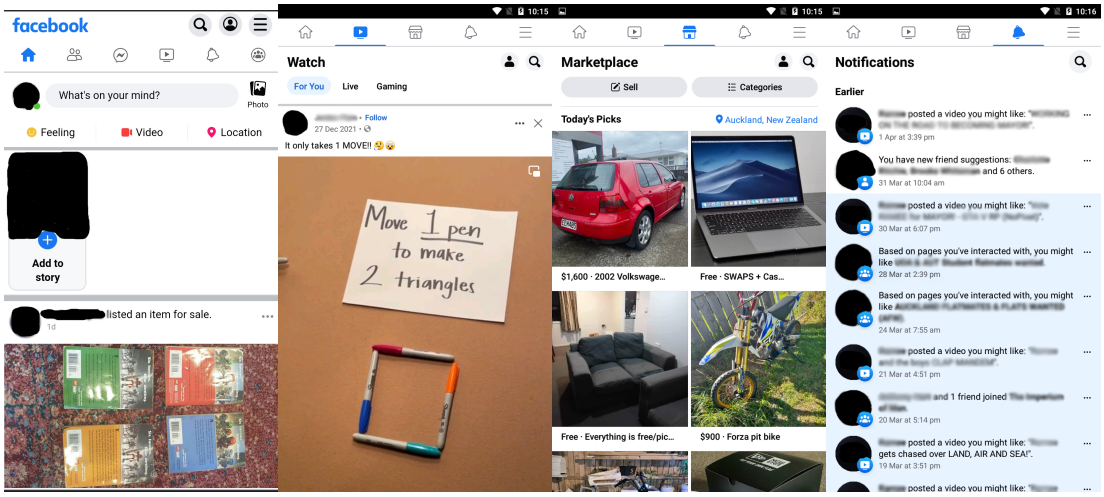
- **Home**: The feed from user friends and following groups, see Figure 5.1a.
- **Watch**: Personalised Videos from friends, following groups and TV shows[2], see Figure 5.1b.
- **Marketplace**[3]: A place people trade things, see Figure 5.1c.
- **Notification**: Show friend suggestion, friends likes and recommended stories, see Figure 5.1d.

EdgeRank

Facebook previously used EdgeRank algorithm for the news feed. EdgeRank has three components which are:

- User affinity score: The relationship between the user and the content/users content, such as commenting, liking and tagging, which generates a score based on those features.

[1] https://www.facebook.com/formedia/tools/news-feed
[2] Launch in 2017
[3] Originally launched in 2007 (closed in 2014), later relaunched in 2016



(a) Home (b) Watch (c) Marketplace (d) Notification

Figure 5.1: Facebook

- Content/Edge weights: The weight of each type of edge (feature/content), e.g., commenting has a higher weight than liking content. Thus higher weight contents are shown more often in the news feed.
- Time Decay: New content holds more value than older posts. Therefore, new posts are shown more.

A simplified version of the EdgeRank[4] algorithm was presented as:

$$\sum_{edge\ e} u_e w_e d_e$$

where

- u_e is user affinity.
- w_e is how the content is weighted.
- d_e is a time-based decay parameter.

[4]https://en.wikipedia.org/wiki/EdgeRank

Edge rank[5] is more sophisticated and optimised than what is shown publicly. The public EdgeRank inner workings are a business-oriented presentation that shows a basic understanding of the algorithm but does not release detailed work. The formula states that if the user interacts with a specific user, e.g., by commenting and liking, they will appear more often or at the top of the news feed. This can also be done with group posts, events, or news stories and topics (Birkbak & Carlsen, 2016).

Machine learning ranking algorithm

In 2017, Facebook used a machine learning ranking algorithm. One research publication into the feed algorithm is feature selection using a group-sparsity-regularised training algorithm for the Facebook feed ranking system (Ni et al., 2019). The regular news feed ranking system has four stages[6]:

1. Inventory, which collects stories from friends and followed topics.
2. Signals contain who posted it when it was posted, reactions (likes, comments), and current internet connections.
3. Prediction, selecting features using an algorithm to generate probability to share, like and comment on the post.
4. The relevancy score is generated by the selected post using a ranking algorithm.

Ni et al. (2019) proposed a method that focuses on feature selection (step 3). The algorithm is called Group Lasso Follow The Regularised Leader. This method adds a group sparsity regularizer (insert method working) to the optimizer in Follow The Regularised Leader (FTRL), allowing for zero feature weight. The research aims to develop an efficient feature selection method for better scalability without losing the

[5]http://edgerank.net/

[6]https://about.fb.com/news/2018/05/inside-feed-news-feed-ranking

performance of the original method. The experiment consists of a Facebook News Feed dataset that contains many feeds with features to extract. The experiment used the proposed method and FTRL to predict users liking of posts using a two 4-layer, fully connected NN. The results showed that their proposed method performed better than the traditional Gradient-Boosted Decision Trees (GBDT) feature selection method. GBDT generates multiple decision trees one at a time (feed-forward learner) and combines the results of each decision tree after each iteration. The result of the feature importance shows that GL features all (no feature has zero weight) while Group Follow The Regularized Leader (G-FTRL) features 5.12%. Next, they tested the algorithm's selected features for the AUC and NE metrics. The results of AUC and NE show similar or better performance with significantly fewer features, thus improving scalability and efficiency.

In 2018, Facebook experimented with not using their news feed algorithm in .05% of users on their platform[7]. The results showed that users would spend more time on Facebook to find interesting stories, thus increasing ad viewing and hiding post requests. Group content was displayed more frequently because friends would comment on group posts that the user did not follow (uninteresting to the user).

Facebook blog

Facebook shares articles on what has changed in the algorithm relating to current affairs, technology improvement and tests. Major changes include:

- In 2013, Facebook improved its new feed[8] by showing older stories that users had not seen before that may interest them. The results showed an increase in likes, comments, and shares by 5% and a 70% increase in stories read.

[7]https://bigtechnology.substack.com/p/facebook-removed-the-news-feed-algorithm?s=r
[8]https://www.facebook.com/business/news/News-Feed-FYI-A-Window-Into-News-Feed

- In 2014, Facebook introduced Trending[9], which shows current affairs topics to discuss. Topics are shown to the user when they follow a topic or a topic that their friends follow or like. Facebook introduced more ways to control[10] their news feed, such as quick access to unfollow or refollow, not interested requests, and reports.

- In 2016, Facebook added a reaction system[11] to react (Wow, Angry, Sad, Love, and Haha) to stories. This system will affect the type of content displayed in the user's news feed depending on the stories' reactions. Facebook also reduced clickbait stories[12] in users' news feeds by introducing an algorithm that identifies clickbait headlines by their common patterns and features and checking links to the headlines. Additionally, in 2017, they improved the algorithm[13] by using it in other languages, checking whether the headline tells the information or exaggerates it, and video clickbait.

- In 2019, Facebook introduced a way to see why stories are shown to the user's news feed[14] called "Why am I seeing this post?". This allows the user to understand why the stories appear on their news feed.

- In 2021, Facebook removed "transparent authorship as a signal to rank news content[15]". Transparent authorship is a system that reduces the likelihood of news articles that do not have a published editorial staff list (not creditable information) being shown on the user's feed. This was removed because of the lack of effect

[9]https://about.fb.com/news/2014/02/news-feed-fyi-showing-stories-about-topics-you-like

[10]https://about.fb.com/news/2014/11/news-feed-fyi-more-ways-to-control-what-you-see-in-your-news-feed/

[11]https://about.fb.com/news/2016/02/news-feed-fyi-what-the-reactions-launch-means-for-news-feed/

[12]https://about.fb.com/news/2016/08/news-feed-fyi-further-reducing-clickbait-in-feed/,https://about.fb.com/news/2017/05/news-feed-fyi-new-updates-to-reduce-clickbait-headlines/

[13]https://about.fb.com/news/2017/08/news-feed-fyi-taking-action-against-video-clickbait

[14]https://about.fb.com/news/2019/03/why-am-i-seeing-this/

[15]https://about.fb.com/news/2020/06/prioritizing-original-news-reporting-on-facebook/

it provided. Furthermore, Facebook added more control to the news feed by introducing filters and favourites[16]. The filter allows access to the algorithm feed or the newest feed, and friends in favourites are ranked higher in the algorithm.

Facebook changes and public outage

Some significant changes were made because of public outrage or privacy concerns. In 2016, Facebook[17] was criticised for misinformation on the 2016 US presidential election[18]. In response to the criticism Facebook reduced clickbait, as explained in the major news feed update section. Another response to the criticism was in 2018 when Facebook introduced Keyword Snooze[19]. Keyword Snooze allows users to mute keywords or phrases in stories from other users (pages, groups, and friends). Most changes occur because of user feedback directly from the app, surveys, privacy concerns from governments, and regulations/lawsuits.

5.1.2 Instagram Home and Suggestion Feed Algorithm

Instagram home feed[20] is a core feature of Instagram where an algorithm finds posts (media with comments, likes, and captions) that the user would be interested/care about from a list of accounts that the user follows. In addition, suggested content is displayed via the user's interactions with followed account's posts in the suggestion feed when the user views all the posts the algorithm finds.

As of 2022, Instagram has the following main features:

- **Home**: following feed and suggested post, see Figure 5.2a.
- **Search**: Users search for posted images, see Figure 5.2b.

[16]https://about.fb.com/news/2021/03/more-control-and-context-in-news-feed/
[17]https://www.facebook.com/zuck/posts/10103269806149061
[18]https://www.businessinsider.com.au/facebook-settlement-ftc-billion-privacy-2019-7
[19]https://about.fb.com/news/2018/06/keyword-snooze-a-new-way-to-help-control-your-news-feed/
[20]https://help.instagram.com/1986234648360433

- **Reels**[21]: Short videos (Up to 60 seconds), see Figure 5.2c.
- **Shop**[22]: Business's sell their products, see Figure 5.2d.

(a) Home (b) Search (c) Reels (d) Shop

Figure 5.2: Instagram

Instagram[23] was the solution for people wanting to share photos and socialise on their mobile phones because there was no other solution.

In 2016[24], Instagram started using a machine learning algorithm from a reverse chronological order feed. In a 2018 reporting conference, the Instagram algorithm's inner workings were revealed. The feed is determined by these factors (signals):

1. Interest in the type of post, e.g., visually, tags that the user interacts.
2. The recency of the post.
3. Relationship of the user and the user's post, e.g., the user tagging, commenting and liking.

[21] Launched in 2020
[22] Tab launched in 2020
[23] https://www.youtube.com/watch?v=a1BBP2Vz0Uc
[24] https://techcrunch.com/2018/06/01/how-instagram-feed-works

Moreover, how Instagram suggest posts is determined below[25]:

1. The user activity: Signals(following, likes, comments).
2. User connection: Connecting to other/similar accounts.
3. Post information: Interactions of the posts.
4. Account information: Interactions to the account.

Instagram introduced the Signals Platform[26], which takes Many to Many relationship pros and centralises them. Signals are metadata that tells us that a photo has an object, e.g., does the photo contain birds, people, or food? This tells Instagram what signals users like in photos to show more personalised posts.

Figure 5.3 is an overview of the relationship between signal producers and consumers and the role of the signals platform.

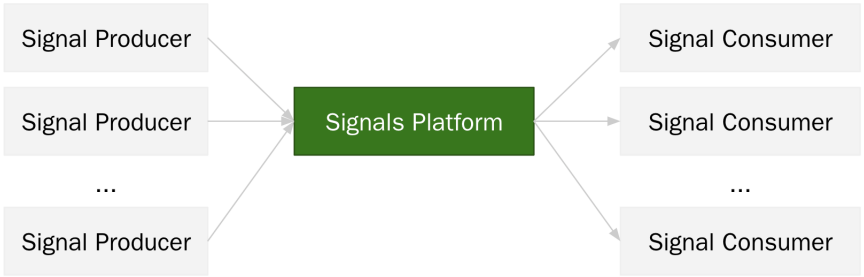


Figure 5.3: Signals platform
Signal Consumer is the algorithm input and Signal Producer is the detector

Below are the signal platform areas:

1. Flow Initiation/Backfilling Signals delay the calculation or calculate the signal at a different time to provide an entry point (signal) to use a key to the flow or a recently uploaded media.

[25]https://help.instagram.com/381638392275939
[26]https://about.instagram.com/blog/engineering/designing-a-signals-platform-for-instagram

2. Model Cascades, machine learning output relies on other models' output, making initiating the flow initiation component challenging. Thus they created an interface to wait for a callback from other models.

3. Signal Post Processing, after the flow initiation is complete, the signal is sent to storage or "pushed to a real-time stream", depending on the user's requests. In Figure 5.4, the user uploads a photo that will be processed through multiple algorithm methods and signals platforms until it reaches post-processing and surface.

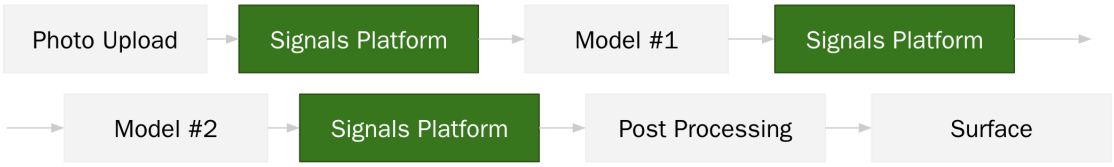


Figure 5.4: Signals flow

A way to overcome scalability issues (an increase in user traffic and a complex method) is to migrate the post-processing storage from bespoke storage to cross-team storage (shared storage). This solution has up-to-date infrastructure, is reliable, and is thus efficient and scalable.

Instagram introduced the Suggest Post system in 2020[27] to allow users to explore other accounts' posts that interest them. After the user views all of their feed, Instagram suggests relevant posts to the user. This system uses:

- K-nearest neighbours (KNN) Embeddings-based similarity, which uses account data to find similar accounts to a seed ("An author or media that one has shown explicit interest towards").

- Co-occurrence-based similarity, which finds patterns to create a list of media that the user interacts with and the frequency of the media pair similar to association

[27] https://about.instagram.com/blog/engineering/designing-a-constrained-exploration-system

rule mining.

From these methods, it generates the top-N list from a seed. The Instagram ranking is determined by the user's engagement with posts and media via reactions such as liking, commenting, and saving, generating value. The weight value is determined by "offline replay over user sessions" and Bayesian optimisation. The choices of models are MTML (Multi-Task Multi-Label Spare Neural Nets), GBDT, and LambdaRank. Instagram solves the cold start problem (not enough user data) in two ways, fallback graph exploration (gathering data from liked user accounts as seeds) and popular media (showing popular media for users to create a profile preference).

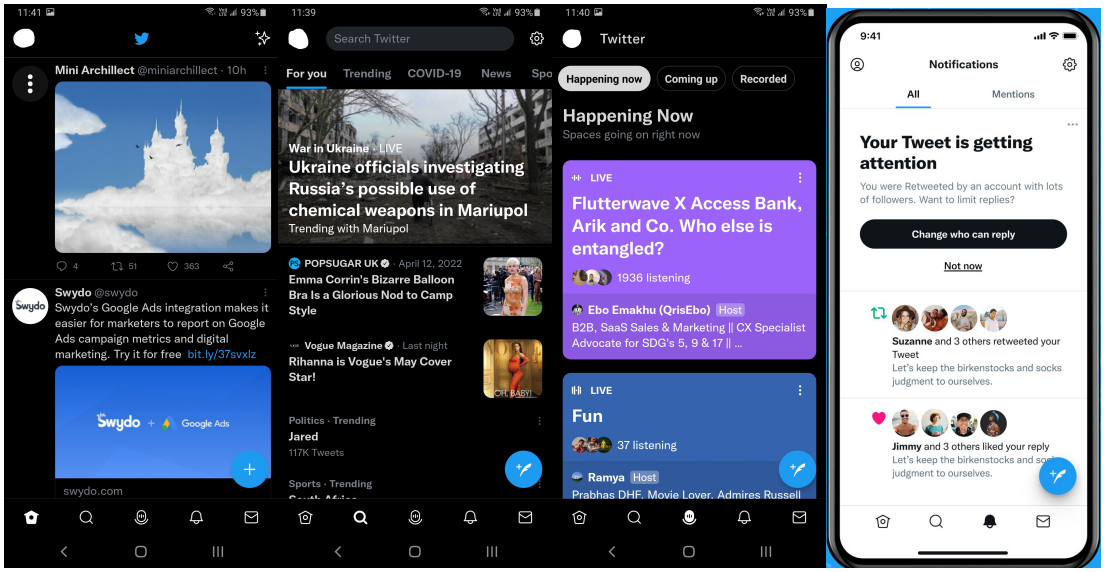
5.1.3 Twitter Timeline Algorithm

A Twitter timeline[28] is a feed that shows personalised tweets from people whom the user follows. The timeline can be personalised or shown in reverse chronological order and shows follows, liked tweets, retweets, and tweets. Additionally, the timeline shows the other users followed accounts in the user followed list that the algorithm recommends following.

As of 2022, Twitter has the following main features:

- **Timeline**: Personalised tweets and topics from user followings, see Figure 5.5a.
- **Search**: For you, Trending hashtags, COVID-19, News, Sports and Entertainment tweets, see Figure 5.5b.
- **Spaces**[29]: Spaces (live voice conversation), see Figure 5.5c.
- **Notification**: liked tweets, retweets and comments from users, see Figure 5.5d.

[28] https://help.twitter.com/en/using-twitter/twitter-timeline
[29] Launched in 2020



(a) Timeline (b) Search (c) Spaces (d) Notification

Figure 5.5: Twitter

Similar to all other algorithms in previous sections, *Using deep learning at scale in Twitter's timelines* (n.d.) states that the algorithm gathers all tweets from the user's followers, creating a list. The algorithm is used to rank which tweets are engaging for the user. The algorithm gives each tweet a score based on its recency, media types, reaction, author's relationship with the user (interactions), past tweet engagement, and Twitter usage. These features combine to create a score in which the highest scored tweets are shown at the top of the timeline. Additionally, when the algorithm runs out of tweets, the "in case you missed it" module will show tweets ranked by their score.

Deep learning is introduced in the timeline, but because of the sparseness of the Twitter data, it is different from the area in which deep learning excels. Moreover, availability, latency, and missing features would have difficulties in achieving Twitter's goal. To overcome the problem, the team at Twitter experimented with the following modifications: Discretize sparse feature inputs, adding two extra layers of sparse layer (online normalisation scheme), adding a "custom isotonic calibration layer to recalibrate and output actual probabilities" (*Using deep learning at scale in Twitter's timelines*,

n.d., p. 1), and training the algorithm in sequence. These improvements gained high accuracy in the offline model and higher tweet and platform engagement in the online model (*Using deep learning at scale in Twitter's timelines*, n.d.). The life cycle of the home timeline[30] is shown below and in the Table 5.6

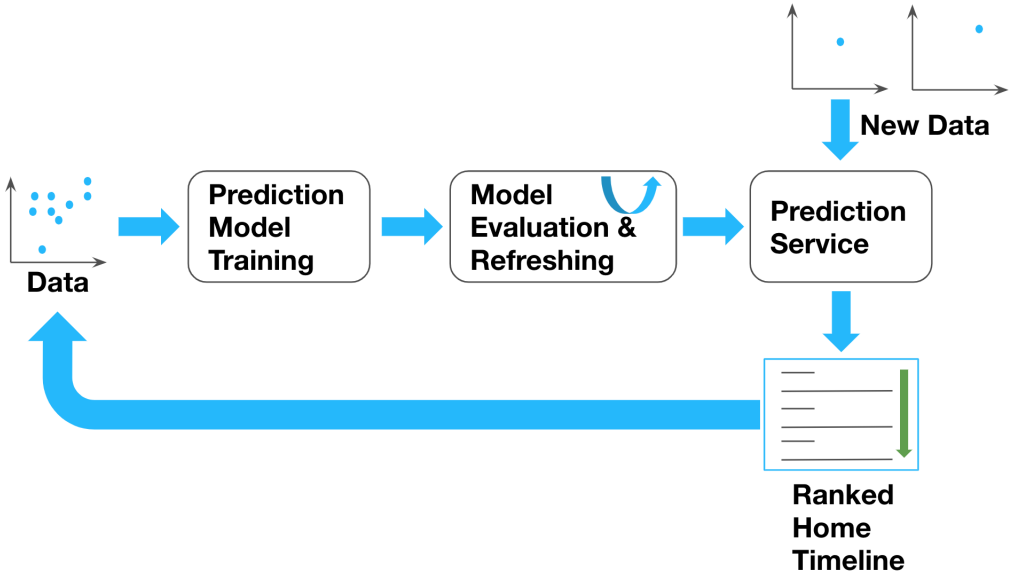


Figure 5.6: Twitter timeline life cycle

1. Data contains user engagements such as retweets, likes, favourites, and server logs as features and labels. Additionally, the features and labels are combined, e.g., tweet A with feature A is labelled positive (engaged) or negative (not engaged) by user A.

   Moreover, the data is downsampled because of imbalanced labels (significantly more negative than positive ones), producing poor results.

   It takes 2-4 days to complete.

2. Prediction Model training is where the processed data is sent to the model to

[30] https://blog.twitter.com/engineering/en_us/topics/infrastructure/2020/streaming-logging-pipeline-of-home-timeline-prediction-system

generate a new model to test (unseen data) its performance. It takes one day to complete.

3. Model Evaluation and Refreshing is performed by evaluating the existing and new models. If the new model performs significantly better than the existing model, it is replaced. It takes one day to complete.

4. Production service, the new model is used publicly in the home timeline.

Producing new models takes significant time. To reduce the time[31], Twitter adopted the Kappa streaming architecture (process two jobs in parallel) to have real-time data processing (training data). The data processing is a Kafka Cluster where user data is split into two streams: Served Keys Topic contains "tweets/user pairs", and Served Features Topic contains "features". In addition, the labelled topic contains user engagements. The next step is to join the Labels Topic with the Served Keys Topic, where the data is downsampled and later joined with the Served Features Topic. The result of the joining is split into multiple engagement types, which are copied to the Hadoop Distributed File System (HDFS). The improvements saw the data processing time decrease from 2-4 days to 4-6 hours and the overall process from 4-6 days to 1 day. Engineers have more time to experiment with new ideas/models, and users on Twitter would have better tweet suggestions from new models.

In 2022, research on solving sparse features was published called graph machine learning with missing node features[32], the objective of the study is to handle missing features from accounts for a Graph Neural Network (GNN) model. This is a problem because accounts would not have all the required features to train, which may decrease accuracy. To solve this problem, Rossi et al. (2021) used Feature Propagation (FP) to fill

[31]https://blog.twitter.com/engineering/en_us/topics/infrastructure/2020/streaming-logging-pipeline-of-home-timeline-prediction-system

[32]https://blog.twitter.com/engineering/en_us/topics/insights/2022/graph-machine-learning-with-missing-node-features

in the missing features. This is done by propagating through features on a graph before sending the data to the GNN model. The experiment dataset consists of Cora, Citeseer, PubMed, Amazon-Computers, Amazon Photo, and OGBN-Arxiv. The performance of the tests (0-99% missing feature) shows high accuracy/similar accuracy as the other models (GCNMF, PaGNN, MGCNN), what FP excels in after 50-70% missing features and loses on average of 2.50% (90% missing features) and an average of 4.12% (99% missing features). Moreover, the efficiency is tested against PaGNN and GCNMF, which showed a significant decrease in runtime as the dataset grew and was up to 3x faster. Rossi et al. (2021) found there are some limitations, such as that it is designed for "graphs with only one node and edge type" (p. 9) and FP "treats feature channels independently" (p. 9) and some ethical concerns if it is used misleadingly, e.g., creating a profile from interest and interaction without the user's consent or submission.

Twitter suggests who to follow[33] that may interest the user. The previous "who to follow" algorithm was based on collaborative filtering and graph expansion using signals (features), which have a limited source of user information. A model-based framework with a two-tower model (generating user behaviour embedding and generation production behaviour embedding) is used to improve the "who to follow" section. The model allows Twitter to input user features (interests, language, favourites) to calculate the user embedding and find the k-nearest embedding in the embedding space to recommend accounts. The combination of the two towers produced a more relevant "who to follow", reduced maintenance, improved efficiency and scalability, and made it easy to add new signals.

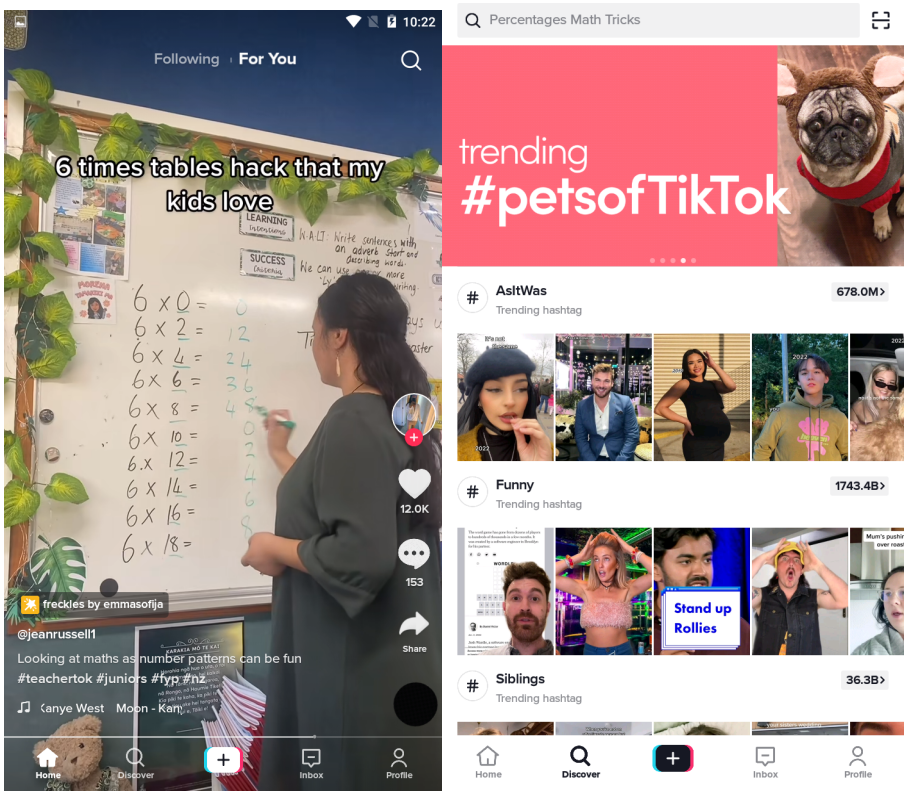
[33]https://blog.twitter.com/engineering/en_us/topics/insights/2022/model-based-candidate-generation-for-account-recommendations

5.1.4 TikTok Algorithm

TikTok[34] is a short video platform to upload, discover, and share with people. The video can be up to 10 minutes, and it consists of the latest trends, dancing, video filtering, and effects. Users can comment, react, and remix or recreate videos from other users.

As of 2022, TikTok has the following main features:

- **Home**: Two types of personalised feeds: *Following*, videos from accounts a user follow *For You*, videos recommended by the service, See Figure 5.7a.
- **Discover**: See trending and search for hashtag video, see Figure 5.7b.



(a) Home (b) Discover

Figure 5.7: TikTok

[34]https://www.tiktok.com/

TikTok For You feed is based on user preference. User preference can be determined by user interactions (liked videos, followed accounts, content created), video information (hashtags, sounds, captions), and account settings (language, performance). The view length and country of creation can indicate the weight of these user interests. The algorithm diversifies the feed by not displaying duplicated content in a row, such as the same song used, duplicate content, and the same creator. TikTok also gets more information about the content by using natural language processing, object detection, and sound detection to get detailed information about what users like to see and hear (Klug, Qin, Evans & Kaufman, 2021; *How Tiktok recommends videos #ForYou*, 2020).

New user preferences are made at the start by selecting their interests from various interests, such as travel and technology. Over time, the interaction of the app creates better recommendations and recommends new interests (*How Tiktok recommends videos #ForYou*, 2020).

Klug et al. (2021) conducted a qualitative interview and quantitative data collection on the TikTok algorithm. The quantitative data collection is done by collecting metadata from videos on TikTok that are trending via the TikTok API. The metadata contains hashtags, view count, share count, comment count, duration, and post time. A qualitative interview is done by questioning 28 participants aged 18-25, 65% female, 300 - 450k followers, 15-450 content creators, and 10k-11.5m likes. Klug et al. (2021) found out that:

- Video engagement, participants assume that the algorithms help the video trend, which is influenced by the high engagement through comments, likes, and shares. The results of the data show that there is an upward correlation.
- Posting time, participants assume that there is a strategy to posting at a high trending time (make a video trend or have high engagement). The data shows that the top 10% of videos were viewed between 6 am and 4 pm UTC, and using

the two-sample Kolmogorov-Smirnov test, they found out that the distribution of the p-value is $1.779e^{-41}$ (smaller than the p-value of 0.05). Therefore, the time posted has a strong influence on whether the video trends or not.

- Hashtags, participates assume that adding trending hashtags or using many hashtags would increase the engagement of the video (Tiktok suggests hashtags when creating the post), and that using #fyp or #foryou would display on users For You pages. The results showed that there is no correlation between having more hashtags and getting more engagements. Furthermore, popular hashtags are not the only contributors to more engagements.

- Delay, they found that there is a delay in TikTok trending hashtags and trending page as trending videos were up to two weeks old and top10% videos were older than the rest.

The study concluded that creators understand what makes their content have better engagement, backed by the trending data collected from TikTok, but adding many hashtags does not correlate with high engagement. Furthermore, the study found that there is a loop where users react "to the behavior of the algorithm, and the algorithm reacts to the behavior of users" (Klug et al., 2021, p. 90) which means that the algorithm learns from the reactions of the users. Therefore, a type of content that creates high engagement would influence what creators post to increase their post engagement.

5.1.5 Algorithm Comparison

Table 5.1 is a summary of the four services. All services post videos and uses comments, likes, and shares to engage with the post. Most services have their goals, but they do not state the more fundamental reason (produce more engagement and revenue). Twitter[38]

[38]https://about.twitter.com/en/who-we-are

Services	Facebook	Instagram	Twitter	TikTok
Post	Story (text, image, video)	Post (image, video)	Tweet (text, image, video)	Post (video)
Engagement	Comment*, Like**, Share***, Reactions	Comment*, Like**, Share***	*,**,***, Retweet	*,**,***, Remix
Features	Home (Stories, Room), Marketplace, Watch (Reels), Notifications (Friend Suggestion, Interest)	Home (Posts), Suggestion Feed	Timeline (Tweets), Search (For you, Trending, COVID-19, News, Sports and Entertainment), Spaces, Notification	Home (For You, Following), Discover
Machine Learning	Ranking of Stories	Ranking of Images	Ranking of Tweets	Ranking of Videos
Goal of the algorithm	Personalised stories from friends and groups.	Personalised posts from followings and relevant posts from other users	Personalised posts from following, recommending who to follow and followings likes	Show trending and Personalised posts.
Goal of the company	More engagement*, Revenue**, "Give people the power to build community and bring the world closer together[35]"	*,**	*,**, "open service that's home to a world of diverse people, perspectives, ideas, and information[36]"	*,**, "to inspire creativity and bring joy[37]"
Key Inputs	Stories (relevancy, recency), User relationships*, User activity/engagement**	Posts (image features, recency),*,**	Tweets (media type, recency), *,**	Post (video features, recency, song), *,**, User information (device performance, location)
Scalability technique	Group Lasso Follow The Regularized Leader (FTRL)	Converted the post-processing storage to cross-team storage (shared storage)	Kappa streaming architecture and Feature Propagation	NA

Table 5.1: Social media services overview

on the other hand, does publicly say about the revenue they make and data from the users and what they do with them. All algorithms use information about the post and user interactions (activity and engagement), but TikTok uses more information by gathering device information and detecting what is in the video using object detection, natural language processing, and sound detection. Instagram also detects objects in their images to create tags/metatags on the post.

5.2 Effects of Social Media Algorithms

In this section, we explore the four effects of social media algorithms, including misinformation, censorship, bias and addiction.

5.2.1 Misinformation

Misinformation is one of the main effects of social media algorithms because of "feedback loops" or bubbles (algorithms provide content based on our likes and beliefs) (Fernández et al., 2021). This is related to Chapter 2 and further explained there. Vosoughi et al. (2018) found that on Twitter, 70% of false information tweets were more likely to be retweeted than true information, and true information takes six times longer to reach the same number of users (1500 people). Low-credible contents tend to be misinformation; popular ones are hard to distinguish from true or false information and are more likely to spread (Shao et al., 2018). Humans are the main contributors to the spread of false information rather than a bot, although a bot makes it easier for the spread to start/reach a wider audience (Vosoughi et al., 2018; Shao et al., 2018; Fernández et al., 2021). Fernández et al. (2021) did an analysis on the amplification of misinformation from recommendation algorithms and found that on a small scale on Twitter (2,921 users), the collaborative filtering method that uses the popular behaviour or matrix factorisation algorithm behaviour showed a high spread of misinformation

(100% and 99.5%) when each user had a 20% misinformation content as a baseline. Nearest neighbour algorithm behaviour (on users) had a higher spread of misinformation than random behaviour (32.7% vs 2.7%). Therefore, popular tweets tend to be misinformation and become popular quickly. The least amount of spread is from using random behaviour because of its nature to ignore the relation/interaction between users and tweets and another called nearest neighbour algorithm behaviour on tweets.

5.2.2 Censorship

Censorship prevents information from being seen or shared by the public, which can be against the freedom of speech. One censorship problem can occur when censoring misinformation (is it against freedom of speech?). Patty (2019) states that free speech is important because it allows the freedom of ideas to be explored by expressing one view on specific topics and sharing them across communities. Suppressing content can shape who we are, control our everyday lives, and make users unable to explore cultures (Wong, 2019; Cobbe, 2020). Governmental officials threaten social media companies to address topics such as fake news on their platforms. However, some content is suitable to be censored, as given by the Supreme Court (Patty, 2019). Social media profit is driven by shareholders and advertisers, which can influence how the service is run and what content it shows (Wong, 2019; Cobbe, 2020). YouTube has been used to promote political advertisements because millions of users would see them. Additionally, YouTube is used to debate issues and express one's opinion on such topics (Patty, 2019). The Communications Decency Act (CDA) states that services such as Google and Facebook do not need to censor topics as this does not hinder free speech (Heins, 2013).

This section explores censorship detection, actions of social media services (Facebook, Google and Youtube), effects and resistance from users.

Detecting content

Facebook maintains content on its platform through manual review or a system. Manuel review involves content reviewers deciding based on the community guidelines (previously private) provided by Facebook (Hooker, 2019). Additionally, reviewers show concerns about handling controversial content not shown in the guidelines (Patty, 2019; West, 2018). Facebook uses an algorithm to show users interesting stories and censor content that they are not interested in, or that breaks the service's terms. Thus, users stay on the platform to see similar content (Wong, 2019).

YouTube has an algorithm that analyses all uploads on its platform to analyse the content and check if it does not break the terms of service. The algorithm is imperfect. Thus, content can be restricted by accident (Patty, 2019; Cobbe, 2020).

Actions of social media

Facebook banned hate speech, violence, graphics, and pornography on their sites as written in their terms of service, and this, however, can impact sex education and the visual arts (Heins, 2013). Moreover, content about religious groups and ethnicities can be subjective as employees of Facebook decide the manual review (Heins, 2013; Cobbe, 2020). Some content is requested to be removed, such as human rights violations, the New Zealand Christchurch terrorist attack, ISIS recruitment, and white supremacy (Patty, 2019; Wong, 2019). Since 2018, Facebook has removed 99.5% of terrorist content, 96% of nudity and sexual content, 86% of violence, and 38% of hate speech automatically (Cobbe, 2020).

Google dominates the search engine sites and censors "offensive" content (determined by Google). Thus, users would not know what is censored (Heins, 2013). People who own the company have a significant say in the company's decisions, one such controversial allegation is when the former Twitter CEO suppressed criticism of Barack

Obama in a Q and A session from spreading on Twitter and another is removing a trending topic (Patty, 2019).

With pressure on YouTube from outside sources and advertisers, YouTube put in new guidelines that prevent targeted harassment, hateful content, or specific topics such as firearms, which removes content that involves firearms in videos, links, and the promotion of Nazi propaganda (Patty, 2019; Hooker, 2019).

Effects of censorship

West (2018) reported on the finding on OnlineCensorship.org that users would submit social media content that the service removed. The objective of the report is to know what type of content is removed, what the impact of removing such content is, and what users' opinions say about the moderation of the service. The findings showed:

- Other user reports users post. An example is when a person targets an individual with content to "harass" them or the reporter does not like a particular religion. Another example is that friends having a "banter" would be seen as harmful by other people, which they could report it (Cobbe, 2020).
- Algorithms "shadowban" (content is not viewable publicly without poster knowledge).
- Account suspension stops users from accessing other services that use the suspended account login. An example is when the user's Facebook account is suspended, and the user can no longer log in to Spotify to access music or lose administrator permissions.
- Lose of communication with friends and family when their account is suspended.
- Lose of work, finances, followers, and communities when their account is banned.

- There are no details of the suspension reasons to stop it from happening again, thus confusing the user.
- There are limited human interactions to appeal the suspension or no instruction to appeal. Thus, users would not know how to appeal and would not get their accounts back.

The content (words) of the LGBTQ community is reported to be suppressed from YouTube (Patty, 2019; Cobbe, 2020). Youtube content creators are pressured to follow the "advertiser-friendly content" guidelines or be demonetised by YouTube. Therefore, it suppresses creators' creativity and opinions (Patty, 2019).

Resistance

Cobbe (2020) found two types of censorship resistance, everyday resistance and organised resistance. Everyday resistance means using safe methods to get around the rules and disguise or conceal the content. Some examples are that users in China would change communication services depending on the type of information being talked about, alter images or videos to avoid detection or create or alter the spelling of words. Organised resistance is when the public gets together to form a group resistance to use legal/non-legal methods to resist censorship. Some examples include creating/moving to a new service (decentralising, open source, and encrypted messaging) and taking the service to court (data protection and privacy).

Summary

Censorship has issues that involve free speech and social media services. Censorship can limit the exploration and sharing of ideas, thus controlling our everyday lives. Governments have put pressure on social media services to address the issues of fake news, and therefore it can influence the content shown on their platform. Social media

services use manual and algorithms to censor content that breaks their terms of service, such as hate speech, violence, and explicit content. Because these content can be subjective and the process is imperfect, it can lead to accidental censoring or banning. There has been resistance to circumventing social media services rules by creating or joining alternative services or taking legal actions.

5.2.3 Bias

Bias[39] is an action in which there is an unfair judgement on a subject. In social media, the algorithms can influence the user because they aim to show relevant content that would bring user engagement. Thus it can be exploited by attackers. Trending topics that are promoted can show bias and low credibility. African-American users' tweets are labelled more offensive than others by two times because of bias (Cobbe, 2020).

Algorithm to reduce the bias of news

Waddell (2019) did a questionnaire on users to see if automated-authored news is seen as less biased and creditable than human-authored news. The data is gathered from 612 participants aged 37.83 years, 53% male and 81% "White/Caucasian" followed by 8% "Black/African American". The author experimented by asking the participants to read an article written by a human journalist, an algorithm, or both a journalist and an algorithm. The articles include short current-event news articles from MSNBC, Fox News, and Automated Insights (Natural Language Generation). The participant would then answer the question that would measure the "perceptions of article credibility, perceived bias, and source anthropomorphism (treating non-humans as humans[40])". Four seven-point semantic differential items are used to measure source anthropomorphism,

[39]https://theconversation.com/misinformation-and-biases-infect-social-media-both-intentionally-and-accidentally-97148

[40]https://www.oxfordlearnersdictionaries.com/definition/english/anthropomorphic

four Likert-type items to measure media bias, and three Likert-type items to measure credibility. The results showed:

1. A one-way ANOVA test showed that algorithm-attributed articles decrease perceived media bias.

2. Algorithm attributed articles decrease perceived source anthropomorphism.

3. The perceived bias from algorithm-attributed articles would indirectly affect their credibility. Additionally, the source anthropomorphism from algorithm-attributed articles has a negative indirect effect on its credibility.

4. Articles written by both journalists and algorithms showed a decrease in media bias and higher credibility.

Chen et al. (2021) experimented with political bias on Twitter using neutral bots. The objective of the experiment is to find out what Twitter users are exposed to. This will help in finding the source and connections of biased information. The accounts to gather information are run by bots called drifters, which are algorithmically controlled (social bots) and have unbiased behaviours. Each drifter first follows a different account and is examined after five months (10th July 2019 - 1st December 2019). They observed that:

1. Total followers of each drifter show what type of users follow them. For instance, they found that drifters with extreme sources as initial follow tend to have more followers, while centre drifters tend to have fewer followers (less influential).

2. Followers from centrist, moderate and partisan drifters show more bot-like interactions than left-leaning followers.

3. Partisan followers are clustered more than centrist, and right-leaning has more echo chambers than left-leaning. Additional "drifters find themselves in structural

echo chambers where they are exposed to content with homogeneous political alignment that mirrors their own. (Chen et al., 2021, p. 3)".

4. Drifters first follow from right-leaning were found to have significantly low credibility (15% of sources were low credibility) shown in 5.8.

5. Drifters first follow were right-leaning tend to stay at that political side while left-leaning "drift towards the political center" as they are exposed to conservative content which they themselves will spread.

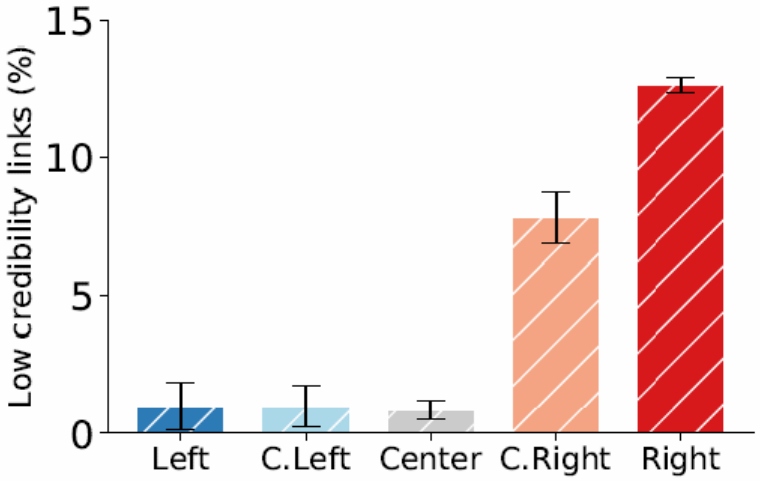


Figure 5.8: Low-credibility links in home timeline
From *Neutral bots probe political bias on social media* (Chen et al., 2021, p. 5)

In summary of Chen et al. (2021) article, when users on Twitter start following right-leaning users, they will be exposed to right-leaning, low-credibility content, and they will spread it as they see more of the content. Twitter users who follow left-leaning users gain more followers and follow more bots, thus having a denser social network. They did not see bias in the Twitter timeline as the drifter was exposed to followings content, but analysing the hashtags shows a slightly central bias on accounts, and links show a significant central bias (liberal and conservative drifters). The experiment

uncovered that Twitter has no bias in its algorithm, but user activities may get users into echo chambers, exposing them to partisan, inauthentic, and misleading content.

Cinelli et al. (2021) studied on echo chamber effect on social media by looking into the "homophily in the interaction networks" (p. 1) and "bias in the information diffusion towards like-minded peers" (p. 1) on Facebook, Twitter, Reddit and Gab. To measure the user's leaning and the spread of information, researchers looked at:

- Twitter, examining the tweet links and user tweets. Three dataset topics of gun control, ObamaCare and abortion.
- Facebook, examining likes stories from users. Three dataset topics of vaccines, science versus conspiracy, and news.
- Reddit, examining links from submissions and comments. Three datasets subreddits from the_donald, politics and news.
- Gab, examining the posts links by users.

Cinelli et al. (2021) found that the algorithms "mediate and influence the content promotion accounting for users' preferences and attitudes" (p. 5). Therefore, Facebook and Twitter are more likely to promote content similar to their leanings (pro or anti), and these services do not provide algorithm options. Moreover, Reddit and Gab are more likely to promote left or right leaning content because the average leaning collected does not equal zero and has adjustable algorithms. In addition, Cinelli et al. (2021) found that Facebook can be seen with significant separation because of the users leanings, while Reddit does not have such separation and users are "more homologous" see Figure 5.9.

5.2.4 Addiction

One of the most common social media problems faced is addiction. D'Arienzo et al. (2019) did a systematic literature review and found that:

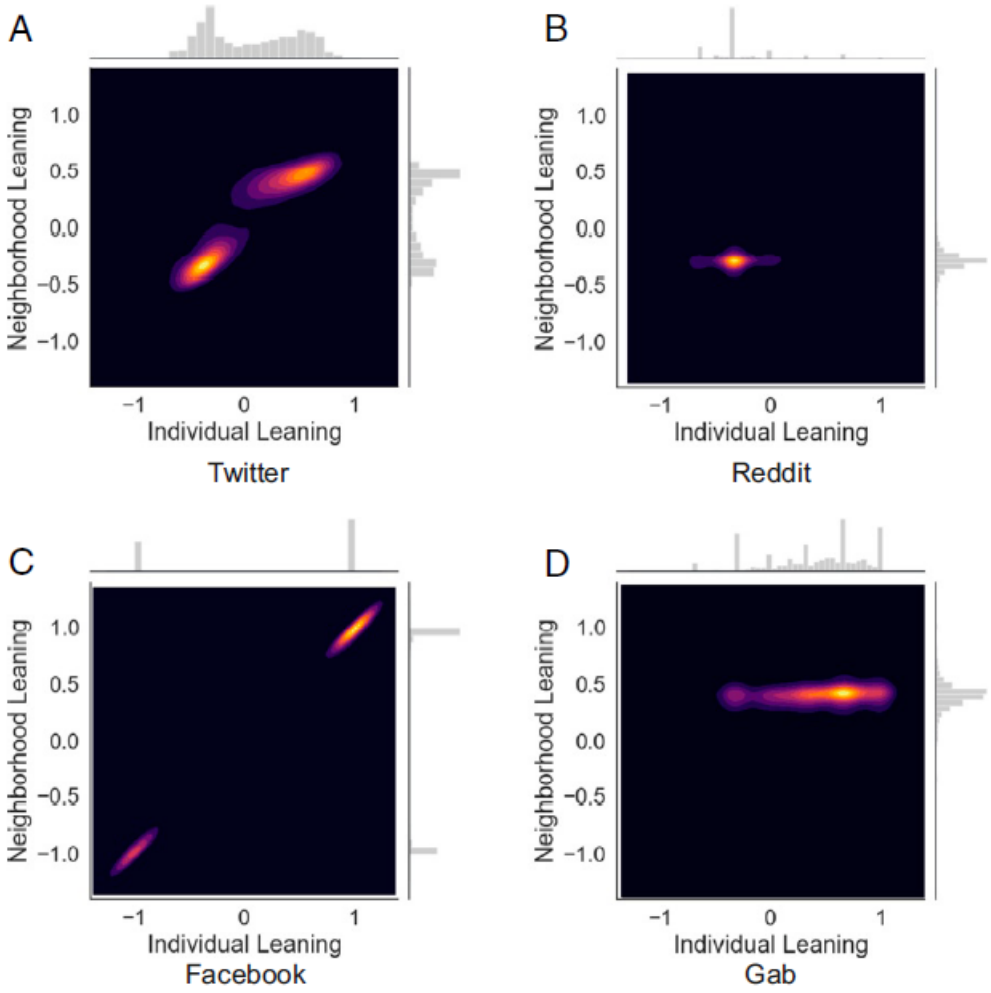


Figure 5.9: Joint distribution of the leaning of users x and the average leaning of their neighborhood x^{NN} for different datasets

From *The echo chamber effect on social media* (Cinelli et al., 2021, p. 3)

- Higher attachment anxiety increases Facebook and SNS use because individuals are more likely to want to start an online relationship (high attachment anxiety likely decreases perceived interpersonal competency). Moreover, users use Facebook late at night.

- Young users who are "preoccupied with their" (p. 1110) offline relationships may use online as a way to escape from reality to solve their loneliness and offline interaction worries.

- Individuals use Facebook to fulfil their need for affection and social interactions because of the loneliness in their lives. Thus, the longer they spend on Facebook, the more interactions occur.

- Users with anxious attachments are more "likely to be associated with general social media activity" (p. 1112) such as commenting, reacting, and scrolling through their feed. They also want to receive feedback from their posts, as it makes them "feel alive". Therefore, they try to create an ideal self-image and personality.

- Young users who have an attachment to their parents are less likely to use the internet to escape their problems (e.g., child abuse), as parental attachment fulfils their needs.

D'Arienzo et al. (2019) concluded that these issues allow one to understand the reasons for the behaviour and what can be done to prevent such issues. D'Arienzo et al. (2019) also found limitations in the previous studies in that the studies relied on self-report rather than seeing the behaviours, as there could be bias.

Aksoy (2018) did a qualitative study on reasons for addictions that 25 students participated in (12 Males and 13 Females). They found that:

1. 24 student used social media because of the lack of friends.

2. 22 were because of "social activity requirement".

3. 21 were because of the feeling of accomplishing a task.

4. 20 were following current events.

5. 16 were "intermingled with social life".

6. 15 were need of making friends.

7. 10 were nothing to do.

8. 23 would not leave social media, but 2 would if there is a good reason to.

These results showed that the main reason for getting addicted to social media is the lack of friends and, secondly, using it as a means of activity. There was no significant difference between genders, but females tend to use social media to communicate with their friends, while males use social media to make new friends (Aksoy, 2018).

Kırık, Arslan, Çetinkaya and Gül (2015); Bhargava and Velasquez (2020) found that individuals addicted to social media ignore real-life responsibility, affecting their real-life relationships and communication. They state that mood is a reason for social media addiction, such as depression, fear, and poor communication skills. An example is the fear of individuals not gaining acceptance in real life and turning to social media to create their own using the environment. Another example is dissatisfaction with real life (falling out of the family and unfulfilling their fantasies), and they turn to social media to find peace and happiness. Individuals addicted to social media see reduced academic and work performance and an inability to keep up with responsibilities. In Turkey, the Ministry of National Education offers a course for elementary schools called "Media Literacy" with eight units: communication, media, television, radio, newspapers, magazines, and the internet. The goal is to learn about the internet, such as accessing information, reading, chatting, problems with the internet, and preventing abuse using the internet. The Radio and Television Supreme Council used a website to show the risks of using the internet for long periods and to prevent addictions. The Information and Communication Technologies Authority restricts users depending on the clients' service in their internet plan. Health services are a treatment service that uses counselling for individuals through weekly or monthly treatments. Kırık et al. (2015) did a quantitative study on 13-19 year-olds to see the level of addiction.

In Figure 5.10 The majority use the internet for 1-3 hours and less than 1 hour,

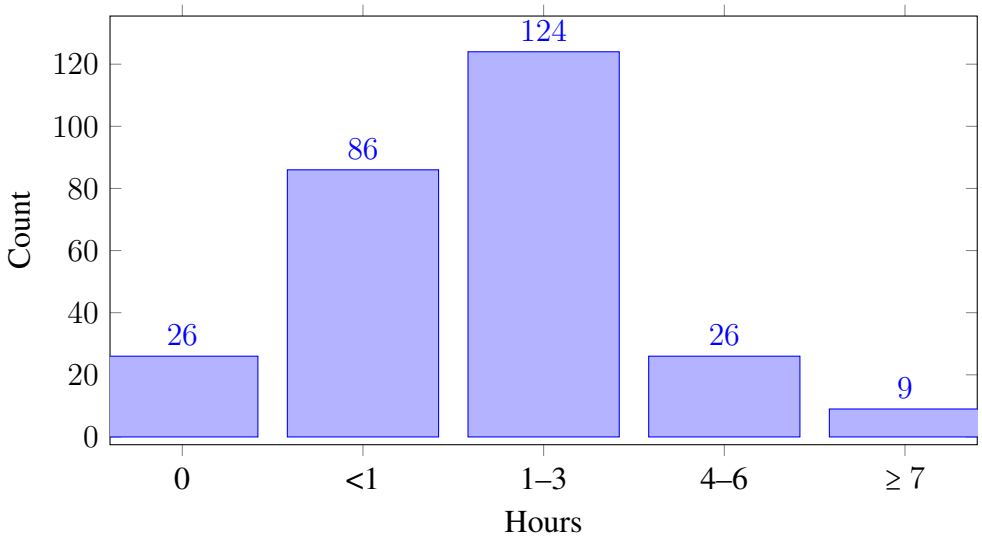


Figure 5.10: Daily time spent on the internet

visiting social media profiles 3-5 times. Moreover, the One Way ANOVA test shows a difference in the level of addiction and frequency of visits to social media profiles. Therefore, the level of addiction increases the frequency of the participant visiting other people's social media profiles. The study's findings showed that early learning about social media addiction and government contributions may prevent addiction.

Nazire Burcin Hamutoglu and Gezgin (2020) studied the direct and indirect effects of social media addiction on social media usage and traits in individuals. The study was performed on 845 students (aged 18-26) from the Sakarya University Faculty of Education in Turkey. The study conducted the examination by using structural equation models. The data was collected using SPSS 23 (statistical software) and AMOS 23 (structural equation modelling). The objective is to analyse the relationship between fear of missing out (FOMO), social media addiction (SMA), and social media usage (DSMU) and the personality traits of extraversion (EXT), agreeableness (AGR), conscientiousness (CON), nervousness (NER) and openness to experience (OE). The analysis showed that SMA and AGR had a direct effect on FOMO. Therefore, increased social media usage increases the FOMO level because of the rewarding experience of

checking on social media.

Aydin et al. (2021) conducted an investigation on social media addiction in adults with depression. They collected 486 participants online (Google forms) over the age of 18 in Turkey using the Social Media Addiction Scale (SMAS) and Beck Depression Inventory (BDI) measurement. The data consist of 419 valid participants, of which 70.4% were women, ages between 18-62 (average 28.46), and 32.2% were students. 99.5% use social media, 52% use it for 1-3 hours, and 8.1% use it for more than 7 hours. Instagram (50.4%) was the most prefered social media, followed by Facebook (23.9%) and 35.1% use it for getting information, 25.8% for leisure time, and 32.5% for entertainment. 39.9% of the participants had minimal depression, 29.8% had mild depression, 21.5% had moderate depression, and 8.8% had severe depression. The results showed that in terms of busyness, emotion, repetition, conflict sub scales and social media addiction, there were no significant differences between males and females. The number of children has a weak relationship with SMAS. Between the ages of 18 and 25, job seekers and students have the highest social media usage. It is seen that social media addiction and depression increase social media usage.

Bhargava and Velasquez (2020) state that social media enables the user's addiction, such as infinite scrolling (no cue to stop scrolling because of never-ending posts), providing rewards (likes) and pulling down for refresh (similar to a slot machine). They exploit the users by not providing default ways to stop the user from using their platform via cues of no new posts or time limit and additionally easier access to delete their account.

5.3 Discussion and Conclusion

In this chapter, we explored the different algorithms used by social media companies and compared them with each other. We found that a significant amount of information

is gathered, from device information to object detection in images. Most companies keep their algorithms' inner workings private, such as Tiktok, and some will tell more than others, such as Instagram. In addition, we tracked the main features that were added are related to the recommendation algorithm and User Interface. Due to the pressure from society and governments, some social media services, such as Facebook, have made changes to reduce false information and provide transparency by showing users the reasons for recommending posts.

We reviewed four main topics related to social effects:

- Misinformation: Misinformation spreads more rapidly than true information because of algorithms' nature to recommend posts that users like from similar users. Thus, social media services need a method to detect and remove misinformation while not hindering freedom of speech (censorship).

- Censorship: If restrictive, it reduces exploration and controls our thinking and voice. Censorship can reduce user harm, although it is up to companies whether the content should be censored. Thus, there is a need to regulate what social media services can or can not censor certain content.

- Bias: Studies show that political leaning depends on the type of content users get recommended, such as false information and echo chambers and depends on the type of social media service they provide. Twitter and Facebook recommend similar leanings content, while Reddit and Gab recommend either one. Thus, similar to misinformation, social media services need to detect biased content and add a footnote to say that the content contains biased opinions.

- Addiction: It profoundly affects young users due to loneliness and as a means of escaping real-life problems to create happiness. In addition, depression increases social media use. Addiction in people showed reduced academic performance

and an inability to keep up with responsibility. Thus, social media services must enforce methods or reduce the algorithm's ability to show new content based on time (show new content every specified time).

The gap in social media services is a need for transparency towards personal data and their algorithms. For example, current public knowledge about algorithms is low, and most users do not know how much information is analysed and tracked.

Chapter 6

Regulations on Data and AI

This chapter surveys data-related regulations and AI laws. First, we explore regulations on data and AI, including surveillance capitalism, data regulations in EU (GDPR) and NZ (NZ Privacy Act 2020), protecting and controlling personal data and technology companies' privacy services (Google, Microsoft, Amazon and Apple). Secondly, we explore AI regulations in the upcoming EU and NZ regulations and its weakness. Finally, we explore the regulation of harmful content in NZ and their content and rights laws (The Films, Videos and Publications Classification Act 1993 and Bill of Rights Act 1990). AI regulation would help gain society's trust, be ethical and reduce the risk of harm to society.

6.1 Personal Data in Technology Companies

Technology companies utilise personal data and surveillance capitalism, and there have been laws put in place to mitigate the harmfulness of them. Due to the law, technology companies have put in user privacy policies. In this section, we explore surveillance capitalism from the Internet of Things, including Google and Facebook, an overview of GDPR and its issues, the latest NZ Privacy Act, and methods to protect and control

personal data, including companies' privacy policies.

6.1.1 Economic Model: Surveillance Capitalism

Surveillance capitalism is a form of information that is predicted and modified for humans to produce profit and control (Zuboff, 2015; Lyon, 2019; Dencik, Hintz & Cable, 2016; Aho & Duffield, 2020). Information is gathered from the 'Internet of Things' in homes (sensor devices), mobile devices, and wearables. Corporations and governments store personal information, such as that of banks, airlines, insurance companies, and telecom communication. These personal information are analysed and sold to data brokers (in the US) without consent or the consumer's knowledge (Zuboff, 2015). Google looked into fees-for-service but was concerned about its user growth. Therefore, they opted for an advertising model using user data to improve their advertising targeting by analysing and using algorithms (Zuboff, 2015). Google collects data using Google searches, emails, text, communication patterns, movements, purchases, clicks, and spellings. With the rise of surveillance, contracts will allow insurance companies to monitor drivers driving safely to keep the insurance/pay violations, or lenders can disable the payer's car if they fail to pay the monthly payment (Zuboff, 2015). In a Pew Research 2014[1], people in 2025 would expect to be tracked and monitored because the benefits of "convenience, safety and services" outweigh privacy. This was a concern in 2010 because users were unaware of how much information is gathered, how the data is treated and monetized, and how long it is stored (Hoofnagle, King, Li & Turow, 2010). Therefore, youth online lack knowledge rather than an attitude towards privacy. Aho and Duffield (2020) states that the attempt to retaliate against surveillance capitalism is the introduction of the GDPR and China embracing surveillance capitalism by using its technology through the public sector.

[1] https://www.pewresearch.org/internet/2014/03/11/digital-life-in-2025/

Facebook uses 'friends', 'likes' and interactions to capture user preferences to be sold to "data brokers, developers, advertisers, political campaigners, and snake-oil vendors" (Lyon, 2019, p. 66). Facebook's objective is to attract users and bring up engagement on the site to learn more about their interests, their daily lives, and their connections, which may be sold to bidders to gain profit (Lyon, 2019).

Google has faced numerous legal crimes: scanning emails of non-Gmail users and students using education apps, collecting voice communications breaching privacy settings, data bundling, saving search data, tracking smartphone locations and wearable devices, and facial detection and recognition (Zuboff, 2015). Google wants greater power to fulfil its vision to 'know what you want and tell you before you ask the question', previously called Google Now 2012 (now called Google Assistant 2016) (Zuboff, 2015).

6.1.2 Regulations of Data Protection in EU and NZ

General Data Protection Regulation (GDPR)

The GDPR is a law on data and privacy protection in EU, which came into effect in 2018.

Data protection principles:

- Lawfulness, fairness and transparency when processing information.
- Must have a valid reason for processing the information.
- Collect and store minimum amount of information and duration as necessary.
- Must keep accurate and up to date information.
- Must process information in a secure protocol.
- Data controller is responsible for following the GDPR law.

GDPR requires employees with access to personal data to use two-factor authentication and storage providers to have end-to-end encryption. In addition, train employees on data privacy policies or limit data access. When an agency has a data breach, they must notify the individual within 72 hours or face penalties (which may be waived if the data is encrypted or useless).

Processing data can only be allowed when the individual gives unambiguous consent to the data processing. It is necessary to use the data to create a contract (background check), comply with a legal obligation (order from court), save somebody's life, perform a task in the public interest or an official function and have a legitimate interest (flexible).

The consent is required:

- to be "freely given, specific, informed and unambiguous".
- to be "clearly distinguishable from the other matters, in an intelligible and easily accessible form, using clear and plain language".
- for individuals to be able to withdraw "previously given consent whenever they want, and you have to honor their decision".
- children under 13 can give consent from parents.
- to keep documented evidence.

(From General Data Protection Regulation 2016 EU)

Individual in EU has the right:

- **to be informed**: About the agency's identity, the processing and purpose, existing of profiling and its consequences, disclosed to the recipient, and data is transferred to international agencies.
- **of access**: Individuals has easy access to their personal information that has been collected to be "aware of, and verify, the lawfulness of the processing."

- **of rectification and erasure**: Of their personal information and no longer be used in processing when their information is not necessary.
- **to restrict processing**: By temporarily moving the information to another processing system and unavailable to users or temporarily removing published information.
- **to data portability**: Individuals can receive information in a "structured, commonly used machine-readable and interoperable format" (encouraged) from an automated process system.
- **to object**: The processing of individual information if the task is for "public interest or in the exercise of official author it vested in the controller, or on the grounds of legitimate interests of a controller or a third party".

(From General Data Protection Regulation 2016 EU)

Politou, Alepis and Patsakis (2018) found concerns over the right to be forgotten and consent revocation. The problem found in the right to be forgotten is informing all controllers of the individual's request to be erased/forgotten because third parties would have copies. Therefore, the original controllers need to implement a system to track where the individual's information is processed or kept and inform them of the request. Although controllers keep links to the copied information, ensuring the individual information is successfully erased is difficult. Politou et al. (2018) found that consent is stricter because consent needs to be unambiguous, informed, and specific. Moreover, the consent does not specify removing the information from storage.

The right to be forgotten prevents the information from being anonymised. Therefore, researchers lose a significant amount of data and lose freedom of speech as the information can be removed ("conflict between privacy and freedom of speech" (Politou et al., 2018, p. 12)). The loss of freedom of speech was seen when the GDPR law went

into effect. Over 1,000 news websites (current affairs, technology, and e-commerce) were inaccessible, including Tribune Media (Los Angeles Times, Chicago Tribune, New York Daily News, Hartford Courant, Orlando Sentinel, and Baltimore Sun) and Lee Enterprises. Thus, Americans and individuals interested in America could not read articles and information from there (Layton, 2019). Innovation and research are at risk, specifically in AI and Big Data. One such example is a research on Iceland's genome warehouse (the oldest genetic record), therapies for Alzheimer's disease, and breast cancer (Layton, 2019)

NZ Privacy Act 2020

The Privacy 2020 Act replaces the Privacy 1993 Act. The act protects individual privacy by giving individuals privacy rights to their personal information. This applies to New Zealand agencies, overseas agencies (carrying business in New Zealand), and individuals who are not ordinarily residents in New Zealand.

The main key points are:

- **Data processing notification**: An agency collecting personal information must notify the user of the reason for the collection.
- **Data transfers**: An agency A can transfer personal information to a foreign/or NZ based agency B when they have permission from the user.
- **Data protection impact assessment**: Privacy assessment is not a requirement.
- **Data protection officer appointment**: An agency requires a 'privacy officer'. It does not apply to an agency who is an individual who collect and hold personal information.
- **Data breach notification**: As a requirement an agency needs to notify the commissioner and the affected individual when there is a data breach.

- **Data retention**: Agency can not hold personal information indefinitely, e.g., no longer it is necessary. If the agency has a 'purpose' for keeping the information, the agency does not need to delete the information.
- **Children's data**: Agency can only collect personal information if the information collected does not intrude further than personal information (e.g., name, age).
- **Special categories of personal data**: The Health Code controls health information. Employment information must be kept private and used for work employment purposes.

(From Privacy Act 2020)

An individual in New Zealand has the right to be informed when their personal information is being processed by providing the recipient's name and address and the name and address of the agency holding the information. Right to access their information readily, to confirm and see what the agency holds, which can lead to the right to rectification (the information is accurate, up-to-date, and not misleading). The right to erasure and the right to object or opt-out is not available unless involved in a court takedown order.

Penalties:

- Individuals that mislead agency to obtain other personal information.
- Failing to notify data breaches.

6.1.3 Methods to Protect and Control Personal Data

There are plenty of guides to protect and control our personal data. Below are the common practices:

- Check privacy settings such as if the profile is private (friends only), public (anyone can see) (*Protecting your privacy online*, n.d.; *7 Tips to Manage Your*

Identity and Protect Your Privacy Online, n.d.; *Protect your privacy on the internet*, n.d.).

- Wary of competitions, questionnaires and wins as they can use the personal information you entered to be sold off to third parties (*Protecting your privacy online*, n.d.).

- Check if the website you are on contains HTTPS. This is to ensure that the information entered is encrypted. Using HTTPS everywhere would also automatically direct to the encrypted version (*Protecting your privacy online*, n.d.; *How to Protect Your Digital Privacy*, n.d.).

- Regular clean up of your social media, remove/unfollow groups that you no longer interact or view (*How to Protect Your Digital Privacy*, n.d.).

- Blocking ads and trackers will decrease the amount of information collected and disable personalised ads (*How to Protect Your Digital Privacy*, n.d.).

- Use an anonymous search engine such as duckduckgo and searX because they do not collect and share your data (*7 Tips to Manage Your Identity and Protect Your Privacy Online*, n.d.; *How to Protect Your Digital Privacy*, n.d.).

- Use a Virtual Private Network (VPN) to stop the user's IP from being tracked by the user's internet service provider. Still, there is an additional concern that the provider of the VPN is trustworthy and does not collect and sell your data (*How to Protect Your Digital Privacy*, n.d.; *How to protect your privacy online*, n.d.; *7 Tips to Manage Your Identity and Protect Your Privacy Online*, n.d.).

- Read the website's privacy policy as this will explain what information is used for, what can be deleted or changed, and where it is shared to (*Protect your privacy on the internet*, n.d.).

- Minimise the amount of information shared publicly, e.g., events, usernames, account numbers and location. Additionally, share the primary email with trusted people or organisations (*Protect your privacy on the internet*, n.d.).
- Search your name in search engines and request removal or review the account visibility settings (*Protect your privacy on the internet*, n.d.).

An overview of the privacy policy and settings of large tech companies

Google

As of 2022, below is the Googles[2] privacy policy and settings

- Advance encryption (HTTPS, transport layer security) is used in emails and storing photos.
- Threat detection and automatic blocking of websites.
- Easy access to control what data is saved in the user's Google account (search, youtube history).
- Prompts a warning if applications need to access information such as locations and cameras on mobile phones.
- Auto delete of personal information, e.g., location history (activity older than 3, 18 months, no auto-delete). In addition, the ability to delete activities (search, assistance and maps).
- Adding data noise to the user information when converting to anonymous information when used for Google Maps (busyness feature).

[2]https://safety.google/security-privacy/

- Sensitive information is kept private from personalised ads such as health, race, and religion. Additionally, the ability to turn off personalised ads and show the user why the ad is displayed.

- Built-in privacy checkup system, where the user can review privacy options such as Ad settings, Photo settings and Auto data deletion (in a step-by-step process).

Microsoft

As of 2022, below are the Microsoft's[3] privacy policy and settings

- Optional diagnostic data (browsing history, website information) when using Microsoft Edge.

- Built-in windows defender smart screen, which blocks malicious website content and downloads.

- Displays what programs use your devices, such as location, microphone and webcam.

- Windows Hello does not store actual images of face, iris and fingerprint. Additionally, biometric verification data will only stay on the device.

- Ability to turn off advertising ID, which prevents apps from using the user's advertising ID.

Amazon

As of 2022, below are Amazon's[4] privacy policy and settings

- Secure protocols (encryption) used when transmitting personal information.

[3]https://privacy.microsoft.com/en-US/privacy-in-our-products
[4]https://www.amazon.com/gp/help/customer/display.html?nodeId=GX7NJQ4ZB8MHFRNJ

- Disable interest-based ads (may still collect data) or disable 3rd party from using an advertising ID.
- Disable some devices from collecting app/device usage.
- Disable location usage.
- Participates in the EU-US and Swiss-US Privacy Shield Framework (2016), which is a data protection framework to transfer data from EU/Swiss to the US (invalid from 2020, yet to be replaced by Trans-Atlantic Data Privacy Framework).
- Receive notice that personal information might be shared with third parties so the user can opt out of sharing their information.

Apple

As of 2022, below are Apple's[5] privacy policy and settings

- Safari, the built-in browser in Apple on IOS and macOS, has Intelligent Tracking Prevention (hides the user's IP address and stops cross-site tracking) and a Privacy Report (displays a report on what the Intelligent Tracking Prevention has blocked). Fingerprinting Defence prevents sites and ads from forming a 'fingerprint' by disguising the user's device (harder to find unique characteristics).
- Maps use end-to-end encryption to sync map data to other users' devices. Apple does not retain history, searches, or places the users have been on maps. Data sent to Apple servers have random identifiers (Apple ID is not used).
- Photo Application using face recognition and object detection is made on the user's device even when using iCloud (stays on the device, and photos are encrypted). Additionally, third-party apps using photos can only access photos that the

[5]https://www.apple.com/nz/privacy/features/

user has selected (limited photo library access).

- Mail, FaceTime, iMessage and SharePlay use end-to-end encryption and do not keep calls on Apple servers.
- Siri, voice processing and suggestions are done on the device. Additionally, any data sent to Apple servers is encrypted and has random identifiers (resets when Siri or Dictation is turned off). Siri and Dictation interactions have defaulted not to share information to improve the services.
- Users can decide which app can use the exact or approximate location (within 26 square kilometres). Users are notified when an application uses location service in the background.
- Apple ID has a built-in function to limit the amount of information other websites can use to log in (ask for name).
- Applications on the App Store must follow Apple privacy guidelines and report how the application uses user data. Additionally, applications must ask permission to use services on the devices.
- Apple shows the data collected with full transparency on the user's data and privacy page.

6.2 AI Regulations

This section explores AI regulations proposals, including general AI regulation, the upcoming EU (EU AI Act 2021), and NZ laws (2020 Whitepaper). In addition, we survey the weakness and effects of the EU AI Act 2021.

6.2.1 A General AI Regulation Proposal

AI brings less privacy, more crime, and unacceptable competition risks. Sun (2021) asks, "how to define the legal nature, legal relationship and legal subject of this algorithm urgently needs the law to play a more and greater guiding role" (p. 90). Answering these questions would allow algorithms to have an "ethical orientation and public interest without having a conflict with relevant law" (p. 90). Sun (2021) proposed:

1. **People-centered principle**: Making human and safety as priority.

2. **Principle of interpretability**: The designer "should explain the source and reliability of the training data" and publish data uses and interests involved in using such algorithm. This proves that what the algorithm has trained on is safe and trustworthy.

3. **Principle of carrying out filing and reviews**: All algorithms/models should be required to be reviewed and documented by internal or government regulatory departments or third-party agencies. Thus, discover any risks from training data and algorithms.

4. **Principle of error correction and fault tolerance**: New algorithms should have spaces for error correction, specifically in personal rights.

5. **Principle of self-supervision**: Designers and operators should be responsible for going through steps to prevent data pollution, algorithm tampering and operation interference.

Regarding regulation at a legal level, it is not complete and the current laws have gaps. The unknowns of AI make it difficult to follow laws and create them. Therefore, regulating the algorithms' designers and operators (service providers) would be effective (Sun, 2021). Creating AI regulation is a complex and lengthy task, as there is a question

of "how to best to regulate AI" by regulating it specifically or regulating different elements of the AI. Without regulation, it is uneasy for developers and users to trust AI because there is no protection from "inequalities and discrimination" and AI is still a developing technology (Madzou, Costigan & MacDonald, 2020).

6.2.2 EU AI Act 2021

The Europe Artificial intelligence Act is an AI regulation proposed in April 2021. As of 14 June 2023, The European Parliament approved its negotiating position on the proposed Artificial Intelligence Act. And the Parliament will negotiate with the EU Council and the European Commission, in the trilogue process.

This regulation aims to develop trustworthy AI systems by following these objectives (*Artificial intelligence act*, n.d.):

1. AI systems in the EU market are safe and respect the existing EU law.
2. "Ensure legal certainty to facilitate investment and innovation in AI" (p. 3).
3. "Enhance governance and effective enforcement of EU law on fundamental rights and safety requirements applicable to AI systems" (p. 03).
4. "Facilitate the development of a single market for lawful, safe and trustworthy AI applications and prevent market fragmentation" (p. 3).

The paper defines four levels of risk, which will help to categorise different AI systems. Below are the 4 risks[6]:

1. **Unacceptable risk:** AI's that threaten the safety, livelihood and rights are banned from use in the EU, such as social scoring, harmful manipulative AI and "exploits vulnerable groups".

[6]https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

2. **High risk:** AI's that impacts people's safety and rights are strictly assessed before use in the EU, such as systems used as safety component/use in "health and safety harmonisation legislation" (toys, cars, lifts), "bio-metric identification and categorisation of natural persons", "education and vocational training".

3. **Limited risk:** AI's that interact with humans, recognition systems, generate/manipulate images, audio, or videos would be required to be transparent, e.g., inform the user of AI use or disclose that it is AI-generated.

4. **Low or minimal risk:** AI's that have minimum risks, such as spam filters or AI used in games, have free use.

Enforcement

Artificial intelligence act proposal (n.d.) states that High-risk AI requires the risk management system to be "established, implemented, documented and maintained" (p. 46) with a "continuous iterative process run throughout the entire lifecycle of a high-risk AI system" (p. 46). The process requires identifying and analysing present and future risks and estimating and evaluating risks that may emerge when using the AI system. High-risk systems using training, validation, and testing data must follow data governance and management practices. The data governance and management practises involving examining possible biases in the data, identifying data gaps (shortcomings and solutions), making the data relevant and free from errors, making appropriate annotations, labelling, and cleaning the data. The high-risk system requires technical documentation, such as before and during the use of the system. The document is required to demonstrate how the system follows the requirement. Logs of the system operation are required, e.g., use period, input data, "identification of the natural persons involved in the verification of the results" (p. 50), and traceable functions used by the system. Additionally, the system must be transparent by showing how the system obtains

the results, how the user can use it appropriately, the contact details of the provider, the intended purpose, the performance (accuracy, limitation, change of performance), the expected lifetime (maintenance, updates), and human oversight (prevent/minimise the risk of health, safety, and fundamental rights).

Concerns

Artificial intelligence act (n.d.) states that the regulation proposal has some concerns. The high-risk-based approach does not fully protect fundamental rights, and the proposal is not detailed enough to recognise which AI systems would do wrong or harm. Therefore, not enough responsibility is assigned, and risk assessment depends on the service provider's self-assessment. Recommendations from academics include broadening the list of banned AI systems, banning existing AI systems, social scoring, and some biometrics systems, providing more detail on classifying the risks, and adding in the sustainability risk of AI systems.

Effect

When the regulation is enforced, many current AI systems must follow its guidelines. Systems that use ranking algorithms that can cause harmful manipulation need to modify their systems to not display manipulative behaviours. This can also be applied to ads that can manipulate users, especially in elections, and users who are undecided about who to vote for. Companies that use AI systems to put users into categories (high-risk systems) to show personalised ads need to follow the AI Act guidelines.

6.2.3 Artificial Intelligence and Law in NZ

In New Zealand, there are ongoing studies on regulating AI. A 2020 white paper called Reimagining Regulation for the Age of AI: New Zealand Pilot Project aims to co-design

an AI regulation in partnership with the Government of New Zealand. Some of the 80% issues are covered by existing legislation but for a short-term solution. Therefore, there is a need for a long-term solution. Madzou et al. (2020) white paper covers three areas:

1. "obtaining of a social licence for the use of AI through an inclusive national conversation" (p. 3).

2. "the development of in-house understanding of AI to produce well-informed policies" (p. 3).

3. "the effective mitigation of risks associated with AI systems to maximize their benefits" (p. 3).

The project would allow AI to become more trusted, protect society, promote innovation, and be ethical. Two workshops were done in New Zealand in 2019 and San Francisco in 2020 to better understand stakeholder issues. The workshop's focus was on keeping citizens safe from AI. The key takeaways were transparency (tested, decision reasons, privacy, life cycle), trust (respecting community views and gaining social licence), people-centred, support for human rights, and "prioritize accountability, fairness, safety and accessibility" (Madzou et al., 2020, p. 7). The recommendations show how the regulation would work, more thorough testing of algorithms (benefit and risk analysis), and new approaches and tools to "encourage investment and innovation" (Madzou et al., 2020, p. 7) from public support and trust. There has been a global focus on AI regulations, such as IEEE's Ethically Aligned Design, which aims to ensure that AI design is robust, safe, and trustworthy. The algorithmic accountability act (introduced to Congress in February 2022) in the United States requires companies to analyse/conduct assessments of their AI system for bias, inaccuracy, or unfairness. The AI regulation proposal in Canada includes data privacy and rights issues applicable to the development and implementation of AI. There are also concerns about the government's AI system

and the public calls for a risk/benefit assessment of AI systems for the government. The risk/benefit assessment will ensure the public that privacy, humans, and ethics are prioritised when designing an AI system using personal data. Such a tool was called the Privacy, Human Rights, and Ethics Framework (PHRaE). It was used when the Ministry of Social Development developed any new services. Therefore, identifying the risks and solutions is important. Below are some guidelines that are being considered for the risk/benefit assessment framework:

- Specify lines of accountability.
- Specify data requirements and flows.
- Define performance metrics.
- Identify the stakeholders.
- Objective of the AI service.
- AI service impact assessment on Human and civil rights.

This white paper starts the beginning of regulating AI in New Zealand and finding a suitable solution that looks at both sides (users and developers).

6.3 Regulation of Harmful Content

Liddecoat (2019) did a study on regulating harmful content in algorithms and social media in New Zealand. New Zealand does not have regulations on AI, yet the country has a robust communication system, and businesses have started developing AI projects. The attack on the Christchurch mosque was streamed and uploaded on Facebook. The video was taken down in 27 minutes but was reuploaded to other services such as 4chan, 8chan, and Twitter, of which the 1.5 million copies were removed within 24 hours. Not

enough was done, as users across the services saw them in their feeds, causing distress, disgust, and inspiration. The legality of the video is ruled a criminal offence to distribute and possess the content. Still, it is available to experts, reporters, and academics by the chief censor, David Shanks. The Internet Service Providers (ISP) were taking a risk by blocking sites whose content had not been declared objectionable. The result of the attack questioned how the algorithms could display such content in the user's feed and if there were any solutions. The challenge in solving the problem is identifying relevant content, as videos can be edited to evade automatic detection. Therefore, there is a need to find the best classifiers that can accurately block them and are ethical for the user (freedom of expression, censorship, and economics). Moreover, misinformation was spread online, such as about gun law reform. Liddecoat (2019) suggested further research on ethical problems faced with various algorithmic classifiers.

In the following, we will briefly mention two laws: The Films, Videos and Publications Classification Act 1993, and Bill of Rights Act 1990. The former is used in the justification of the ban of the Christchurch terror attack video footage spreading online, as the video depicts and promotes extreme violence and terrorism. The latter is a law that protecting New Zealanders' civil and political rights, including the freedom of expression.

6.3.1 The Films, Videos and Publications Classification Act 1993

This act regulates what people in New Zealand can view or hear, such as "Films, DVDs, music recordings, books, magazines, sound recordings, images, computer games and publications downloaded from the internet[7]" The following are the key points relating to this research:

[7]https://www.cab.org.nz/article/KB00000870

- **Inspector may seize publications**: The inspector may seize the film if they discover a person is offering a banned film.
- **Seizure of objectionable publications**: The inspector can seize the publication when the inspector "believes, on reasonable grounds, to be objectionable."
- **Search warrants for offences against specified sections**: Can issue a search warrant if the objectionable publication is believed to be kept "for the purpose of being so dealt with", an "evidence of the commission of an offence of that kind", "intended to be used for the purpose of committing an offence of that kind."
- **Takedown notices for objectionable online publications (inserted in 2022)**: Inspectors can issue a takedown notice if the online content is "likely to be objectionable" (under section 22A) or classified as objectionable (under section 23) or "Inspector believes, on reasonable ground" that it is objectionable. Inspector can request the host to remove or prevent access from the public before giving a takedown notice.
- **Online content host must comply with takedown notice**: The content host must remove or prevent public access, including all copies (subject of the notice to or have access/control over it). The host must preserve a copy (securely) of the content for an investigation or processing if the notice of takedown requests one.

(From Films, Videos, and Publications Classification Act 1993)

6.3.2 Bill of Rights Act 1990

The Bill of Rights protects all New Zealanders' civil and political rights. The categories are

- **Life and security of the person**: Right not to be deprived of life, the right not to be subjected to torture or cruel treatment, the right not to be subjected to medical

or scientific experimentation (without the person's consent) and the right to refuse medical treatment.

- **Democratic and civil rights**: Electoral rights (New Zealand citizen over 18 years of age has the right to vote), Freedom of thought (includes conscience, religion, belief, hold opinions without interference), Freedom of expression (freedom to seek, receive, impart information and opinions), Manifestation of religion and belief (right manifest person's religion or belief individually or in community with other in publicly or privately), Freedom of peaceful assembly, Freedom of association, Freedom of movement (New Zealand citizen has a right to leave and enter New Zealand, and Everyone has right to leave New Zealand).

- **Non-discrimination and minority rights**: Freedom from discrimination, Rights of minorities (people in New Zealand who has ethnic, religious or linguistic minority has a right to enjoy their culture, practise their religion and use their language).

- **Search, arrest and detention**: Unreasonable search and seizure (everyone has the right to deny unreasonable search or seizure), Liberty of the person (right to not be arbitrarily arrested or detained), Rights of persons arrested or detained (informed of the reason and right to consult and instruct lawyers).

- **Criminal procedure**: Right to a fair and public hearing, right to be presumed innocent until proved guilty, "right not to be a witness or to confess guilt", and more.

- **The right to justice**: Right to apply for judicial review and to bring proceedings against/defend.

(From New Zealand Bill of Rights Act 1990)

6.4 Discussion and Conclusion

This chapter explored different types of laws related to data, content, and AI. New Zealand's regulation on personal data is weak in privacy and protection compared to the EU's GDPR, such as having no regulation on agents using two-factor authentication (2FA) to access customer data and not having a privacy assessment as a requirement. Although GDPR has strict data laws, it has some weaknesses, such as the right to be forgotten (difficult to anonymise data), keeping track of data, and informing third parties that the data owner wants it erased. Therefore, there is a risk that user's personal data will not be fully erased. There is a need to update the NZ privacy act to be on par with GDPR or better, as the current regulation is weaker than GDPR.

Users need to be made aware of companies using information about them. Several court cases, including Google and Facebook, have involved breaching privacy rules and tracking. Although these companies hold information, they do provide some privacy protection. We explored Google, Microsoft, Amazon, and Apple's privacy terms and services. We saw that Amazon has the worst privacy measures compared to the other three. Apple has the best privacy measures because of encryption, information not being kept on servers, and using maps anonymously.

There has yet to be an AI law in New Zealand. There has been a white paper on AI law in New Zealand since 2020, and it aims to understand what is needed in the regulation for long-term law. The EU has proposed an AI Act in 2021 that separates the types of AI into four risks (unacceptable risk, high risk, limited risk and low/minimal risk). Each risk has the type of AI and what needs to be followed. There are concerns over the proposed law, such as needing more detail on what AIs would harm or make mistakes and a longer list of banned AIs. AI laws must be strict on what AI is allowed to be used, tested thoroughly, and documented on how it works and the training data. In addition, it needs to be long-term and updated.

The Christchurch mosque attack showed how harmful content can spread fast via social media services and what can be done to the algorithms to stop showing harmful content or ways to detect them quicker. The Films, Video and Publications Classification Act 1993 added additional laws in 2022 that requests can be issued to content hosts to remove or block access before takedown notices can be issued. There is a need for social media services to remove/ be responsible for harmful content quickly while not hindering freedom of speech.

Chapter 7

Conclusion

In this thesis, we first reviewed false information, its effect, and methods to reduce it. Secondly, we proposed an algorithm to detect false COVID-19 information from tweets using questions to answer. The model (TOKOFOU) was modified to improve multi-language scores in the NLP4IF 2021 workshop (TOKOFOU_T). The training consists of 20 epochs with a learning rate of $3 \cdot 10^{-5}$ and a weight decay of 0.01. The improvements include using multilingual language models and increasing epochs. The result of the improved algorithm performed the best average (2.2% better than the 2nd best) across the three models, although it performed lower than the original algorithm in English by 1.7%. The performance of LLMs shows a great opportunity to utilise these methods in society and would greatly benefit fact-checkers.

Thirdly, we analysed one of the newest LLMs as of 2022. chatGPT performed unexpectedly well without any training in English (1.5% lower than TOKOFOU). We also fine-tuned GPT3 Ada using the training set, which resulted in better performance in some questions (Q2, Q4, Q5) in English. The performance of chatGPT and fined-tuned GPT3 Ada in other languages (Arabic and Bulgarian) showed lower performance than expected. Thirdly, the explainability of TOKOFOU is done using LIME and SHAP. We found that questions 1 and 3 tend to have similar keywords and rankings (around

90% correlation coefficient), and questions 4, 5, and 7 (which tend to have over 50% correlation coefficient) tend to have some correlation to each other. There is a need for more research on explainability because we lack knowledge of the meaning behind the keywords in neural networks.

Lastly, we explored social media services, including Facebook, Instagram, Twitter and TikTok and their algorithm updates. We also explored the algorithm's effects on users (Misinformation, Censorship, Bias and Addiction). In addition, we explore laws that relate to privacy (GDPR, NZ Privacy 2020), AI (EU AI Act 2021 and NZ Whitepaper), personal information in technology companies and its privacy policies (Google, Microsoft, Amazon and Apple). We conclude that users need privacy awareness because they do not necessarily know what social media companies store about them or how they analyse and share it with third parties.

7.1 Limitation

In this thesis, we ran into some limitations. These include datasets, models, and compute time.

Firstly, creating our own dataset that followed the question would have taken time, so we used an existing dataset in the workshop. In addition, further dataset analysis showed human judgement variation in the answer to the question. Therefore, it can affect the outcome and training of the models.

Secondly, instead of training models from scratch, we used pre-trained models and finally fine-tuned them further due to using personal devices. Therefore, there could have been biases in the pre-trained model used in TOKOFOU_T.

Thirdly, Arabic and Bulgarian may have improved performance by asking chatGPT questions in Arabic and Bulgarian instead of English when the tweet either Arabic or Bulgarian. Lastly, using a personal device shows only four examples of explainability

because it is time-consuming for LIME and SHAP to output results.

7.2 Future Work

In the future work of this thesis, we will create an unbiased dataset based on the existing dataset, and further inspection of the ground truth is needed as some of the ground truth answers seen are not agreeable.

Training models or using based models from scratch would reduce bias as there could have been biases in the pre-trained model used in TOKOFOU_T. Therefore affecting the individual scores. In addition, creating a model using tweet links would further validate the answer from the model.

Further research into explainability in the text model is needed, as we had a hard time explaining how TOKOFOU outputs the prediction.

The possibility of using fine-tuned GPT to detect false information by asking if a tweet or article is false or true information quickly and easily. In addition, possibly using GPT to help fact-checkers check the post for false information is efficient.

References

7 tips to manage your identity and protect your privacy online. (n.d.). https://staysafeonline.org/blog/7-tips-to-manage-your-identity/. (Accessed: 2022-05-18)

Acemoglu, D., Ozdaglar, A. & ParandehGheibi, A. (2010). Spread of (mis)information in social networks. *Games and Economic Behavior*, *70*(2), 194-227. Retrieved from https://www.sciencedirect.com/science/article/pii/S0899825610000217 doi: https://doi.org/10.1016/j.geb.2010.01.005

Adadi, A. & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, *6*, 52138-52160. doi: 10.1109/ACCESS.2018.2870052

Aho, B. & Duffield, R. (2020, April). Beyond surveillance capitalism: Privacy, regulation and big data in europe and china. *Economy and Society*, *49*(2), 187–212. Retrieved from https://doi.org/10.1080/03085147.2019.1690275 doi: 10.1080/03085147.2019.1690275

Aksoy, M. E. (2018). A qualitative study on the reasons for social media addiction. *European Journal of Educational Research*, *7*(4), 861-865. Retrieved from https://doi.org/10.12973/eu-jer.7.4.861 doi: 10.12973/eu-jer.7.4.861

Alam, F., Dalvi, F., Shaar, S., Durrani, N., Mubarak, H., Nikolov, A., ... Nakov, P. (2020). *Fighting the covid-19 infodemic in social media: A holistic perspective and a call to arms.* arXiv. Retrieved from https://arxiv.org/abs/2007.07996 doi: 10.48550/ARXIV.2007.07996

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I. & Atkinson, P. M. (2021, July). Explainable artificial intelligence: an analytical review. *WIREs Data Mining and Knowledge Discovery*, *11*(5). Retrieved from https://doi.org/10.1002/widm.1424 doi: 10.1002/widm.1424

Antoun, W., Baly, F. & Hajj, H. (2020). *Arabert: Transformer-based model for arabic language understanding.* arXiv. Retrieved from https://arxiv.org/abs/2003.00104 doi: 10.48550/ARXIV.2003.00104

Arslan, F., Caraballo, J., Jimenez, D. & Li, C. (2020, May). Modeling factual claims with semantic frames. In *Proceedings of the 12th language resources and evaluation conference* (pp. 2511–2520). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/

2020.lrec-1.306

Artificial intelligence act. (n.d.). https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2021)698792. (Accessed: 2022-06-8)

Artificial intelligence act proposal. (n.d.). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206. (Accessed: 2022-06-8)

Aydin, S., Koçak, O., Shaw, T. A., Buber, B., Akpinar, E. Z. & Younis, M. Z. (2021, April). Investigation of the effect of social media addiction on adults with depression. *Healthcare*, *9*(4), 450. Retrieved from https://doi.org/10.3390/healthcare9040450 doi: 10.3390/healthcare9040450

Barbieri, F., Camacho-Collados, J., Neves, L. & Espinosa-Anke, L. (2020). *Tweeteval: Unified benchmark and comparative evaluation for tweet classification.* arXiv. Retrieved from https://arxiv.org/abs/2010.12421 doi: 10.48550/ARXIV.2010.12421

Bert multilingual base model (cased). (n.d.). https://huggingface.co/bert-base-multilingual-cased. (Accessed: 2020-10-10)

Bhargava, V. R. & Velasquez, M. (2020, October). Ethics of the attention economy: The problem of social media addiction. *Business Ethics Quarterly*, *31*(3), 321–359. Retrieved from https://doi.org/10.1017/beq.2020.32 doi: 10.1017/beq.2020.32

Birkbak, A. & Carlsen, H. (2016, January). *The world of edgerank: Rhetorical justifications of facebook's news feed algorithm.*

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). *Language models are few-shot learners.* arXiv. Retrieved from https://arxiv.org/abs/2005.14165 doi: 10.48550/ARXIV.2005.14165

Chang, Y. K., Literat, I., Price, C., Eisman, J. I., Chapman, A., Gardner, J. & Truss, A. (2020). News literacy education in a polarized political climate: How games can teach youth to spot misinformation. *Harvard Kennedy School Misinformation Review*. doi: 10.37016/mr-2020-020

Chen, W., Pacheco, D., Yang, K.-C. & Menczer, F. (2021, September). Neutral bots probe political bias on social media. *Nature Communications*, *12*(1). Retrieved from https://doi.org/10.1038/s41467-021-25738-6 doi: 10.1038/s41467-021-25738-6

Chiou, L. & Tucker, C. (2018, November). *Fake news and advertising on social media: A study of the anti-vaccination movement* (Working Paper No. 25223). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w25223 doi: 10.3386/w25223

Cho, J.-H., Rager, S., O'Donovan, J., Adali, S. & Horne, B. D. (2019, jun). Uncertainty-based false information propagation in social networks. *Trans. Soc. Comput.*, *2*(2). Retrieved from https://doi.org/10.1145/3311091 doi: 10.1145/3311091

Cinelli, M., Morales, G. D. F., Galeazzi, A., Quattrociocchi, W. & Starnini, M. (2021,

February). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, *118*(9). Retrieved from https://doi.org/10.1073/pnas.2023301118 doi: 10.1073/pnas.2023301118

Cobbe, J. (2020, October). Algorithmic censorship by social platforms: Power and resistance. *Philosophy &; Technology*, *34*(4), 739–766. Retrieved from https://doi.org/10.1007/s13347-020-00429-0 doi: 10.1007/s13347-020-00429-0

Confalonieri, R., Coba, L., Wagner, B. & Besold, T. R. (2020, October). A historical perspective of explainable artificial intelligence. *WIREs Data Mining and Knowledge Discovery*, *11*(1). Retrieved from https://doi.org/10.1002/widm.1391 doi: 10.1002/widm.1391

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... Stoyanov, V. (2019). *Unsupervised cross-lingual representation learning at scale.* arXiv. Retrieved from https://arxiv.org/abs/1911.02116 doi: 10.48550/ARXIV.1911.02116

Criado, J. I., Sandoval-Almazan, R. & Gil-Garcia, J. R. (2013). Government innovation through social media. *Government Information Quarterly*, *30*(4), 319-326. Retrieved from https://www.sciencedirect.com/science/article/pii/S0740624X1300083X doi: https://doi.org/10.1016/j.giq.2013.10.003

D'Arienzo, M. C., Boursier, V. & Griffiths, M. D. (2019). Addiction to social media and attachment styles: A systematic literature review. *International Journal of Mental Health and Addiction*, *17*, 1094-1118.

Das, A. & Rad, P. (2020). *Opportunities and challenges in explainable artificial intelligence (xai): A survey.* arXiv. Retrieved from https://arxiv.org/abs/2006.11371 doi: 10.48550/ARXIV.2006.11371

De keersmaecker, J. & Roets, A. (2017). 'fake news': Incorrect, but hard to correct. the role of cognitive ability on the impact of false information on social impressions. *Intelligence*, *65*, 107-110. Retrieved from https://www.sciencedirect.com/science/article/pii/S0160289617301617 doi: https://doi.org/10.1016/j.intell.2017.10.005

Dencik, L., Hintz, A. & Cable, J. (2016). Towards data justice? the ambiguity of anti-surveillance resistance in political activism. *Big Data & Society*, *3*(2), 2053951716679678. Retrieved from https://doi.org/10.1177/2053951716679678 doi: 10.1177/2053951716679678

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv. Retrieved from https://arxiv.org/abs/1810.04805 doi: 10.48550/ARXIV.1810.04805

Dong, M., Zheng, B., Quoc Viet Hung, N., Su, H. & Li, G. (2019). Multiple rumor source detection with graph convolutional networks. In *Proceedings of the 28th acm international conference on information and knowledge management* (p. 569–578). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3357384.3357994 doi:

10.1145/3357384.3357994

Fenn, K., Griffin, N., Uitvlugt, M. & Ravizza, S. (2014, 05). The effect of twitter exposure on false memory formation. *Psychonomic bulletin & review*, *21*. doi: 10.3758/s13423-014-0639-9

Fernández, M., Bellogín, A. & Cantador, I. (2021). *Analysing the effect of recommendation algorithms on the amplification of misinformation.* arXiv. Retrieved from https://arxiv.org/abs/2103.14748 doi: 10.48550/ARXIV.2103.14748

Ghanem, B., Rosso, P. & Rangel, F. (2020, apr). An emotional analysis of false information in social media and news articles. *ACM Trans. Internet Technol.*, *20*(2). Retrieved from https://doi.org/10.1145/3381750 doi: 10.1145/3381750

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S. & Yang, G.-Z. (2019, December). XAI—explainable artificial intelligence. *Science Robotics*, *4*(37). Retrieved from https://doi.org/10.1126/scirobotics.aay7120 doi: 10.1126/scirobotics.aay7120

Guo, B., Ding, Y., Yao, L., Liang, Y. & Yu, Z. (2020, jul). The future of false information detection on social media: New perspectives and trends. *ACM Comput. Surv.*, *53*(4). Retrieved from https://doi.org/10.1145/3393880 doi: 10.1145/3393880

Hanu, L. & Unitary team. (2020). *Detoxify.* Github. https://github.com/unitaryai/detoxify.

Hassan, N., Zhang, G., Arslan, F., Caraballo, J., Jimenez, D., Gawsane, S., ... Tremayne, M. (2017, aug). Claimbuster: The first-ever end-to-end fact-checking system. *Proc. VLDB Endow.*, *10*(12), 1945–1948. Retrieved from https://doi.org/10.14778/3137765.3137815 doi: 10.14778/3137765.3137815

Heins, M. (2013). The brave new world of social media censorship. *Harv. L. Rev. F.*, *127*, 325. Retrieved from https://harvardlawreview.org/2014/06/the-brave-new-world-of-social-media-censorship/

Hodgin, E. & Kahne, J. (2018, 01). *Misinformation in the information age: What teachers can do to support students.* Retrieved from https://www.researchgate.net/publication/341763454_Misinformation_in_the_Information_Age_What_Teachers_Can_Do_to_Support_Students

Hoofnagle, C. J., King, J., Li, S. & Turow, J. (2010, Apr). How different are young adults from older adults when it comes to information privacy attitudes and policies? *SSRN Electronic Journal*. doi: 10.2139/ssrn.1589864

Hooker, M. P. (2019). Censorship, free speech & facebook: Applying the first amendment to social media platforms via the public function exception. *Wash. j. law technol. arts*, *15*(1), 36.

How tiktok recommends videos #foryou. (2020, Jun). TikTok. Retrieved from https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you

How to protect your digital privacy. (n.d.). https://www.nytimes.com/

guides/privacy-project/how-to-protect-your-digital-privacy. (Accessed: 2022-05-18)

How to protect your privacy online. (n.d.). https://us.norton.com/internetsecurity-privacy-protecting-your-privacy-online.html. (Accessed: 2022-05-18)

Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H. & Habash, N. (2021). *The interplay of variant, size, and task type in arabic pre-trained language models.* arXiv. Retrieved from https://arxiv.org/abs/2103.06678 doi: 10.48550/ARXIV.2103.06678

Ireton, C. & Posetti, J. (2018). *Journalism, 'fake news' & disinformation: Handbook for journalism education and training*. United Nations Educational, Science, and Cultural Organization.

Karduni, A. (2019). Human-misinformation interaction: Understanding the interdisciplinary approach needed to computationally combat false information. *CoRR*, *abs/1903.07136*. Retrieved from http://arxiv.org/abs/1903.07136

Klug, D., Qin, Y., Evans, M. & Kaufman, G. (2021). Trick and please. a mixed-method study on user assumptions about the tiktok algorithm. In *13th acm web science conference 2021* (p. 84–92). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3447535.3462512 doi: 10.1145/3447535.3462512

Kumar, S. & Shah, N. (2018). False information on web and social media: A survey. *CoRR*, *abs/1804.08559*. Retrieved from http://arxiv.org/abs/1804.08559

Kırık, A. M., Arslan, A., Çetinkaya, A. & Gül, M. (2015, 10). A quantitative research on the level of social media addiction amongyoung people in turkey. , *3*, 108-122. doi: 10.14486/IntJSCS444_

Layton, R. (2019). *10 problems of the gdpr: The us can learn from the eu's mistakes and leapfrog its policy*. American Enterprise Institute. Retrieved from https://www.judiciary.senate.gov/download/03/12/2019/layton-testimony

Liddecoat, J. (2019). *Algorithms and social media: A need for regulations to control harmful content?* Retrieved from https://giswatch.org/node/6180

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). *Roberta: A robustly optimized bert pretraining approach.* arXiv. Retrieved from https://arxiv.org/abs/1907.11692 doi: 10.48550/ARXIV.1907.11692

Longo, L., Goebel, R., Lecue, F., Kieseberg, P. & Holzinger, A. (2020). Explainable artificial intelligence: Concepts, applications, research challenges and visions. In *Lecture notes in computer science* (pp. 1–16). Springer International Publishing. Retrieved from https://doi.org/10.1007/978-3-030-57321-8_1 doi: 10.1007/978-3-030-57321-8_1

Lundberg, S. & Lee, S.-I. (2017). *A unified approach to interpreting model predictions.* arXiv. Retrieved from https://arxiv.org/abs/1705.07874 doi: 10.48550/ARXIV.1705.07874

Lyon, D. (2019, 03). Surveillance capitalism, surveillance culture and data politics 1. In (p. 64-77). doi: 10.4324/9781315167305-4

Madzou, L., Costigan, M. & MacDonald, K. (2020, 06). *Reimagining regulation for the age of ai: New zealand pilot project* (White Paper). World Economic Forum.

Mekki, A. E., Mahdaouy, A. E., Berrada, I. & Khoumsi, A. (2022, July). AdaSL: An unsupervised domain adaptation framework for arabic multi-dialectal sequence labeling. *Information Processing &; Management*, *59*(4), 102964. Retrieved from https://doi.org/10.1016/j.ipm.2022.102964 doi: 10.1016/j.ipm.2022.102964

Meske, C., Bunde, E., Schneider, J. & Gersch, M. (2020, December). Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities. *Information Systems Management*, *39*(1), 53–63. Retrieved from https://doi.org/10.1080/10580530.2020.1849465 doi: 10.1080/10580530.2020.1849465

Mishra, S., Prasad, S. & Mishra, S. (2020, May). Multilingual joint fine-tuning of transformer models for identifying trolling, aggression and cyberbullying at TRAC 2020. In *Proceedings of the second workshop on trolling, aggression and cyberbullying* (pp. 120–125). Marseille, France: European Language Resources Association (ELRA). Retrieved from https://aclanthology.org/2020.trac-1.19

Müller, M., Salathé, M. & Kummervold, P. E. (2020). *Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter.* arXiv. Retrieved from https://arxiv.org/abs/2005.07503 doi: 10.48550/ARXIV.2005.07503

Nakov, P., Corney, D., Hasanain, M., Alam, F., Elsayed, T., Barrón-Cedeño, A., ... Martino, G. D. S. (2021). Automated fact-checking for assisting human fact-checkers. Retrieved from https://arxiv.org/abs/2103.07769 doi: 10.48550/ARXIV.2103.07769

Nazire Burcin Hamutoglu, M. T. & Gezgin, D. M. (2020, April). Investigating direct and indirect effects of social media addiction, social media usage and personality traits on FOMO. *International Journal of Progressive Education*, *16*(2), 248–261. Retrieved from https://doi.org/10.29329/ijpe.2020.241.17 doi: 10.29329/ijpe.2020.241.17

Nguyen, D. Q., Vu, T. & Nguyen, A. T. (2020). BERTweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations.* Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/2020.emnlp-demos.2 doi: 10.18653/v1/2020.emnlp-demos.2

Ni, X., Yu, Y., Wu, P., Li, Y., Nie, S., Que, Q. & Chen, C. (2019). Feature selection for facebook feed ranking system via a group-sparsity-regularized training algorithm. In *Proceedings of the 28th acm international conference on information and knowledge management* (p. 2085–2088). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3357384.3358114 doi: 10.1145/3357384.3358114

Nuseir, M. (2018, 09). Impact of misleading/false advertisement to consumer behaviour. international journal of economics and business research, 2018 vol.16 no.4, pp.453 - 465. *International Journal of Economics and Business Research*, *16*, pp.453 - 465. doi: 10.1504/IJEBR.2018.095343

Patty, M. (2019). Social media and censorship: Rethinking state action once again. *Mitchell Hamline LJ Pub. Pol'y & Prac.*, *40*, 99.

Politou, E., Alepis, E. & Patsakis, C. (2018, 03). Forgetting personal data and revoking consent under the GDPR: Challenges and proposed solutions. *Journal of Cybersecurity*, *4*(1). Retrieved from https://doi.org/10.1093/cybsec/tyy001 (tyy001) doi: 10.1093/cybsec/tyy001

Protecting your privacy online. (n.d.). https://cert.govt.nz/individuals/guides/protecting-your-privacy-online/. (Accessed: 2022-05-18)

Protect your privacy on the internet. (n.d.). https://support.microsoft.com/en-us/windows/protect-your-privacy-on-the-internet-ffe36513-e208-7532-6f95-a3b1c8760dfa. (Accessed: 2022-05-26)

Ribeiro, M. T., Singh, S. & Guestrin, C. (2016). *"why should i trust you?": Explaining the predictions of any classifier.* arXiv. Retrieved from https://arxiv.org/abs/1602.04938 doi: 10.48550/ARXIV.1602.04938

Ridcully, M. (2003). *Framenet.* Retrieved from https://framenet.icsi.berkeley.edu/fndrupal/CJFFNintroPPT

Rossi, E., Kenlay, H., Gorinova, M. I., Chamberlain, B. P., Dong, X. & Bronstein, M. M. (2021). On the unreasonable effectiveness of feature propagation in learning on graphs with missing node features. *CoRR*, *abs/2111.12128*. Retrieved from https://arxiv.org/abs/2111.12128

Saez-Trumper, D. (2019). Online disinformation and the role of wikipedia. *CoRR*, *abs/1910.12596*. Retrieved from http://arxiv.org/abs/1910.12596

Safaya, A., Abdullatif, M. & Yuret, D. (2020). KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the fourteenth workshop on semantic evaluation.* International Committee for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/2020.semeval-1.271 doi: 10.18653/v1/2020.semeval-1.271

Samek, W. & Müller, K.-R. (2019). Towards explainable artificial intelligence. In *Explainable AI: Interpreting, explaining and visualizing deep learning* (pp. 5–22). Springer International Publishing. Retrieved from https://doi.org/10.1007/978-3-030-28954-6_1 doi: 10.1007/978-3-030-28954-6_1

Schmidt, T., Salomon, E., Elsweiler, D. & Wolff, C. (2021). Information behavior towards false information and ?fake news? on facebook: The influence of gender, user type and trust in social media. In *Information between data and knowledge* (Vol. 74, pp. 125–154). Glückstadt: Werner Hülsbusch. Retrieved from https://epub.uni-regensburg.de/44942/ (Session 2: Information Behavior and Information Literacy 2)

Seaton, J., Sippitt, A. & Worthy, B. (2020). Fact checking and information in the age of covid. *The Political Quarterly*, *91*(3), 578-584. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-923X.12910 doi: https://doi.org/10.1111/1467-923X.12910

Shaar, S., Alam, F., Da San Martino, G., Nikolov, A., Zaghouani, W., Nakov, P. & Feldman, A. (2021, June). Findings of the NLP4IF-2021 shared tasks on fighting the COVID-19 infodemic and censorship detection. In *Proceedings of the fourth workshop on nlp for internet freedom: Censorship, disinformation, and propaganda* (pp. 82–92). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.nlp4if-1.12 doi: 10.18653/v1/2021.nlp4if-1.12

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A. & Menczer, F. (2018, November). The spread of low-credibility content by social bots. *Nature Communications*, *9*(1). Retrieved from https://doi.org/10.1038/s41467-018-06930-7 doi: 10.1038/s41467-018-06930-7

Skurnik, I., Yoon, C., Park, D. C. & Schwarz, N. (2005, 03). How Warnings about False Claims Become Recommendations. *Journal of Consumer Research*, *31*(4), 713-724. Retrieved from https://doi.org/10.1086/426605 doi: 10.1086/426605

Sun, H. (2021, September). Legal examination and regulation of artificial intelligence algorithm. In *2021 international conference on computer information science and artificial intelligence (CISAI)*. IEEE. Retrieved from https://doi.org/10.1109/cisai54367.2021.00119 doi: 10.1109/cisai54367.2021.00119

Tziafas, G., Kogkalidis, K. & Caselli, T. (2021). *Fighting the covid-19 infodemic with a holistic bert ensemble.* arXiv. Retrieved from https://arxiv.org/abs/2104.05745 doi: 10.48550/ARXIV.2104.05745

Using deep learning at scale in twitter's timelines. (n.d.). Twitter. Retrieved from https://blog.twitter.com/engineering/en_us/topics/insights/2017/using-deep-learning-at-scale-in-twitters-timelines

van der Linden, S. (2022, March). Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature Medicine*, *28*(3), 460–467. Retrieved from https://doi.org/10.1038/s41591-022-01713-6 doi: 10.1038/s41591-022-01713-6

Verdict classifier. (n.d.). https://huggingface.co/saattrupdan/verdict-classifier. (Accessed: 2020-10-10)

Vosoughi, S., Roy, D. & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146-1151. Retrieved from https://www.science.org/doi/abs/10.1126/science.aap9559 doi: 10.1126/science.aap9559

Waddell, T. F. (2019, January). Can an algorithm reduce the perceived bias of news? testing the effect of machine attribution on news readers' evaluations of bias, anthropomorphism, and credibility. *Journalism &; Mass Communication Quarterly*, *96*(1), 82–100. Retrieved from https://doi.org/10.1177/1077699018815891 doi: 10.1177/1077699018815891

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N. & Zhou, M. (2020). *Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.* arXiv. Retrieved from https://arxiv.org/abs/2002.10957 doi: 10.48550/ARXIV.2002.10957

West, S. M. (2018, May). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media &; Society*, *20*(11), 4366–4383. Retrieved from https://doi.org/10.1177/1461444818773059 doi: 10.1177/1461444818773059

Wong, K. (2019). Ideological control and social media.

Xlm roberta base snli mnli anli xnli. (n.d.). https://huggingface.co/symanto/xlm-roberta-base-snli-mnli-anli-xnli. (Accessed: 2020-10-10)

Zannettou, S., Sirivianos, M., Blackburn, J. & Kourtellis, N. (2019, may). The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *J. Data and Information Quality*, *11*(3). Retrieved from https://doi.org/10.1145/3309699 doi: 10.1145/3309699

Zareie, A. & Sakellariou, R. (2021). Minimizing the spread of misinformation in online social networks: A survey. *Journal of Network and Computer Applications*, *186*, 103094. Retrieved from https://www.sciencedirect.com/science/article/pii/S1084804521001168 doi: https://doi.org/10.1016/j.jnca.2021.103094

Zuboff, S. (2015, 03). Big other: Surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology*, *30*. doi: 10.1057/jit.2015.5

Appendix A

Glossary

2FA Two-Factor Authentication.

AI Artificial Intelligence.

API Application Programming Interface.

BERT Bidirectional Encoder Representations from Transformers.

CNN Convolutional Neural Network.

FTRL Follow The Regularised Leader.

GBDT Gradient-Boosted Decision Tree.

GNN Graph Neural Network.

GPT Generative Pre-trained Transformer.

LLM Large Language Model.

LSTM Long Short-Term Memory Network.

LIME Local interpretable model-agnostic explanation.

NLP Natural Language Processing.

NN Neural Network.

RNN Recurrent Neural Network.

SHAP SHapely Additive exPlanations.

GCN Graph Convolutional Network.

XAI Explainable AI.

Appendix B

Additional Keywords

B.1 Introduction

Below are the additional material from explainability. These are the complete output of LIME and SHAP

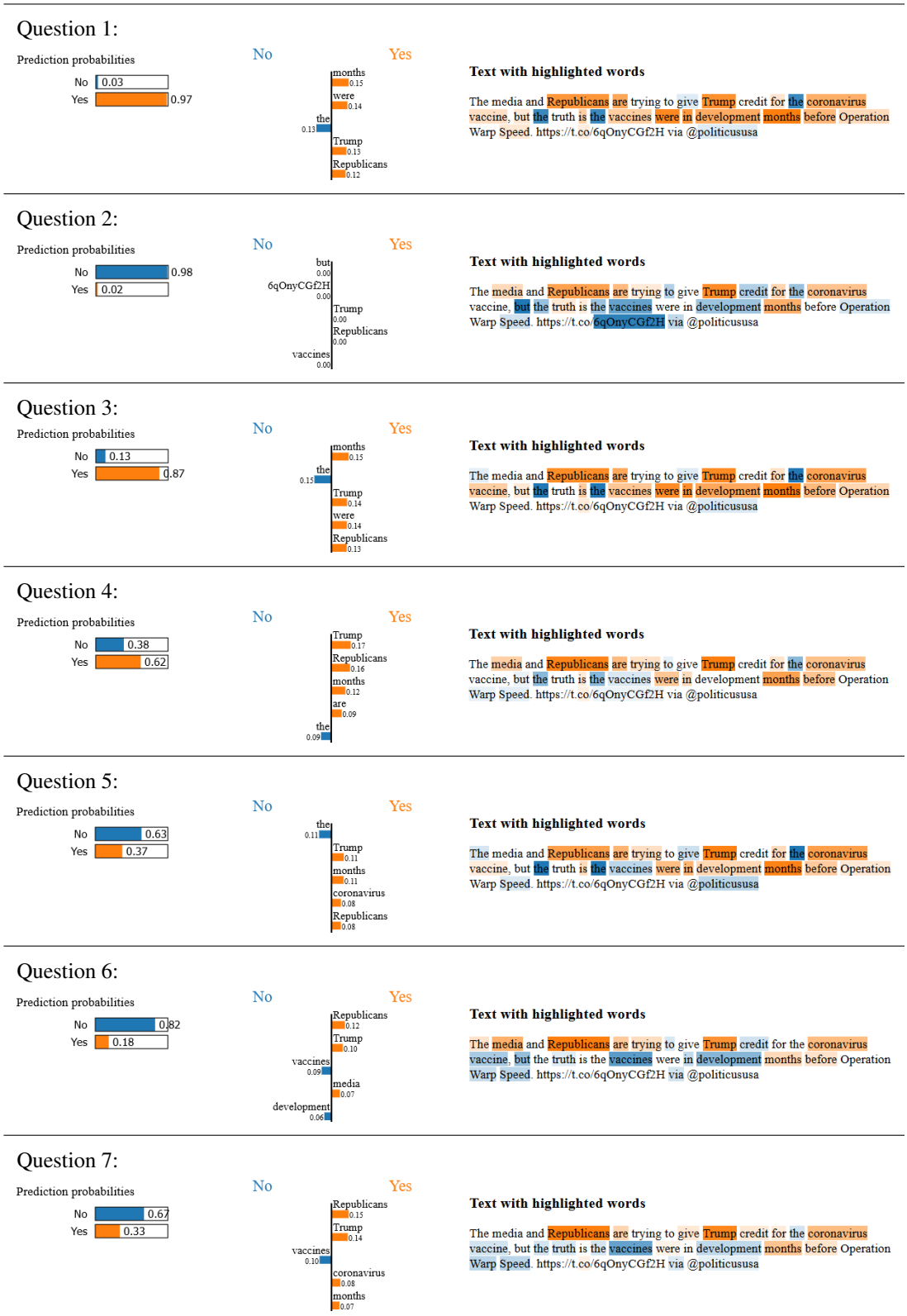


Figure B.1: Tweet ID: 927 LIME Overview

Actual answer: Yes, No, Yes, Yes, No, No, No & Predicted answer: Yes, No, Yes, Yes, No, No, No

Figure B.2: Tweet ID: 1020 LIME Overview

Actual answer: No Nan Nan Nan Nan No No & Predicted answer: Yes No Yes No No No No

Figure B.3: Tweet ID: 1066 LIME Overview

Actual answer: Yes Nan Yes Yes Yes Yes Yes & Predicted answer: Yes No Yes Yes No No No

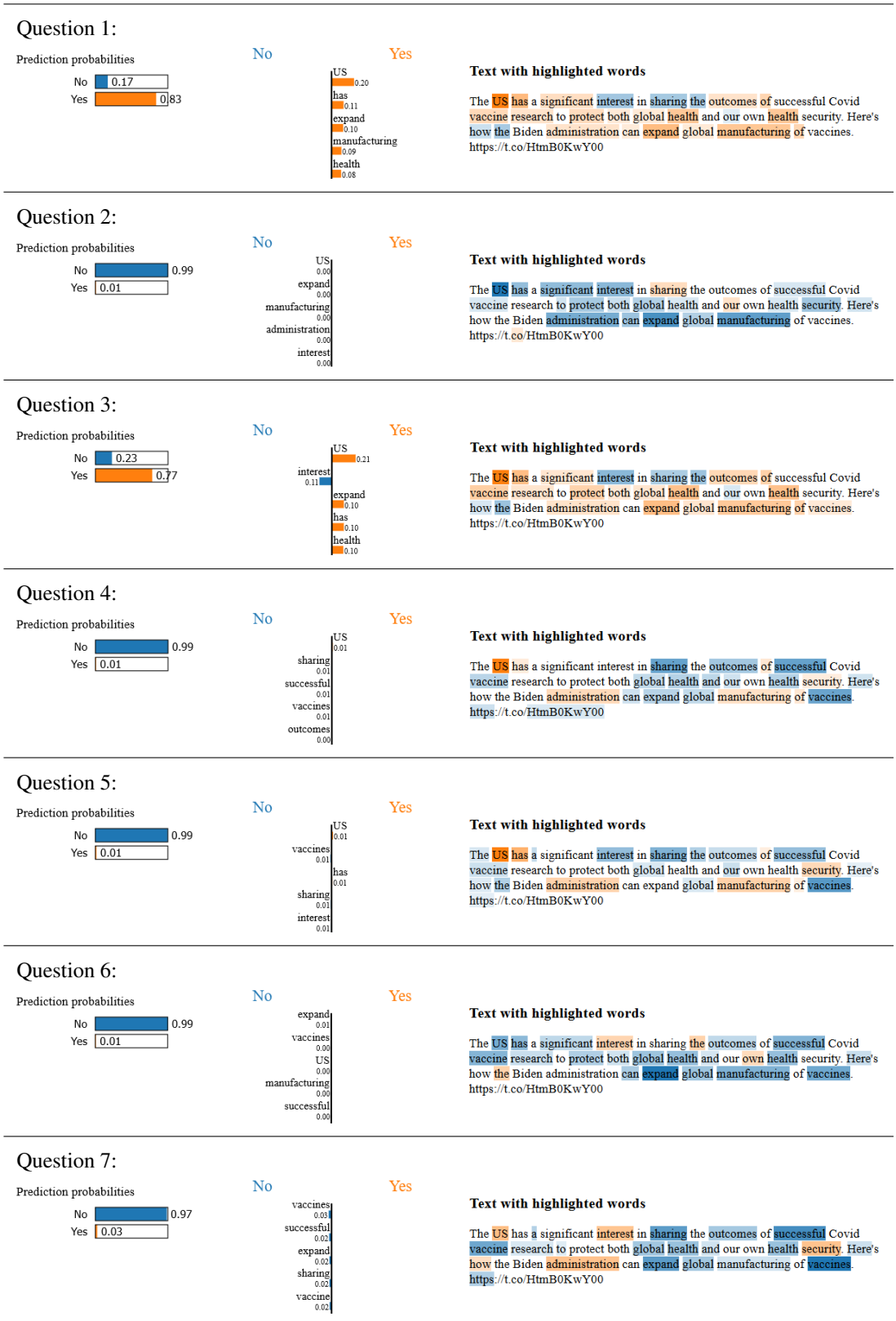


Figure B.4: Tweet ID: 962 LIME Overview

Actual answer: No Nan Nan Nan Nan No No & Predicted answer: Yes No Yes No No No No

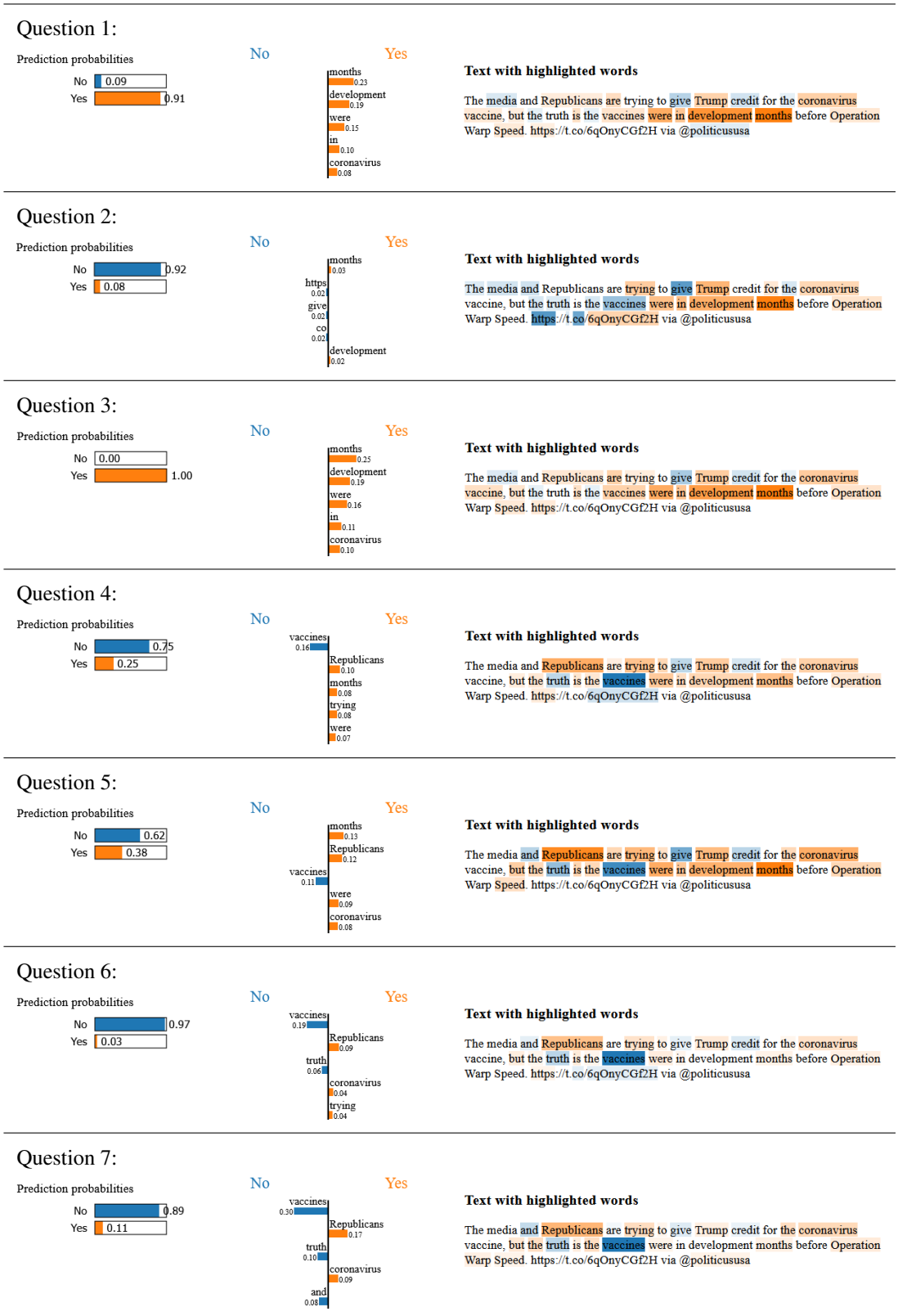


Figure B.5: Tweet ID: 927 LIME Overview for finetune GPT3 ADA

Actual answer: Yes, No, Yes, Yes, No, No, No & Predicted answer: Yes, No, Yes, No, No, No, No

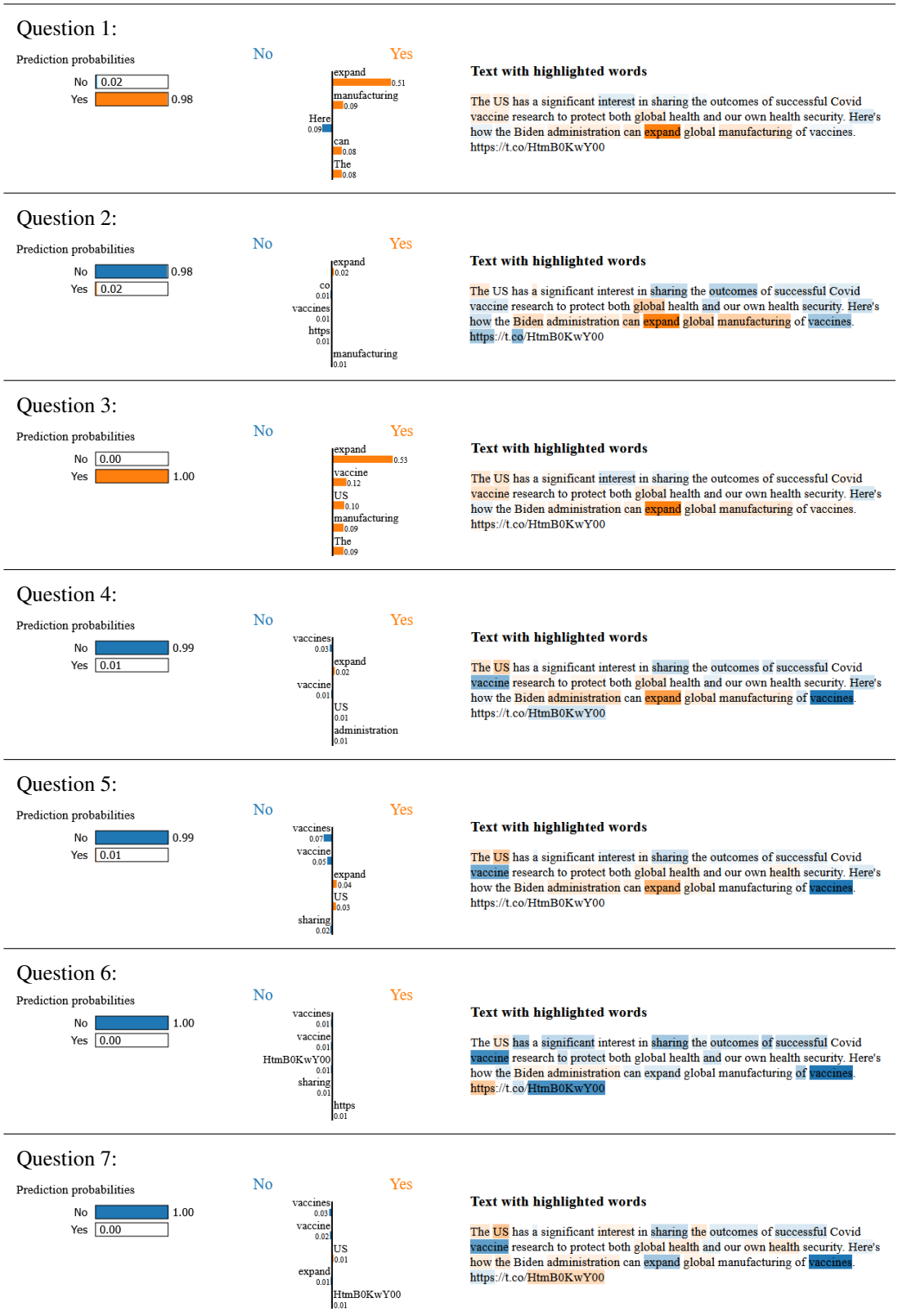


Figure B.6: Tweet ID: 962 LIME Overview for finetune GPT3 ADA

Actual answer: No Nan Nan Nan Nan No No & Predicted answer: Yes No Yes No No No No

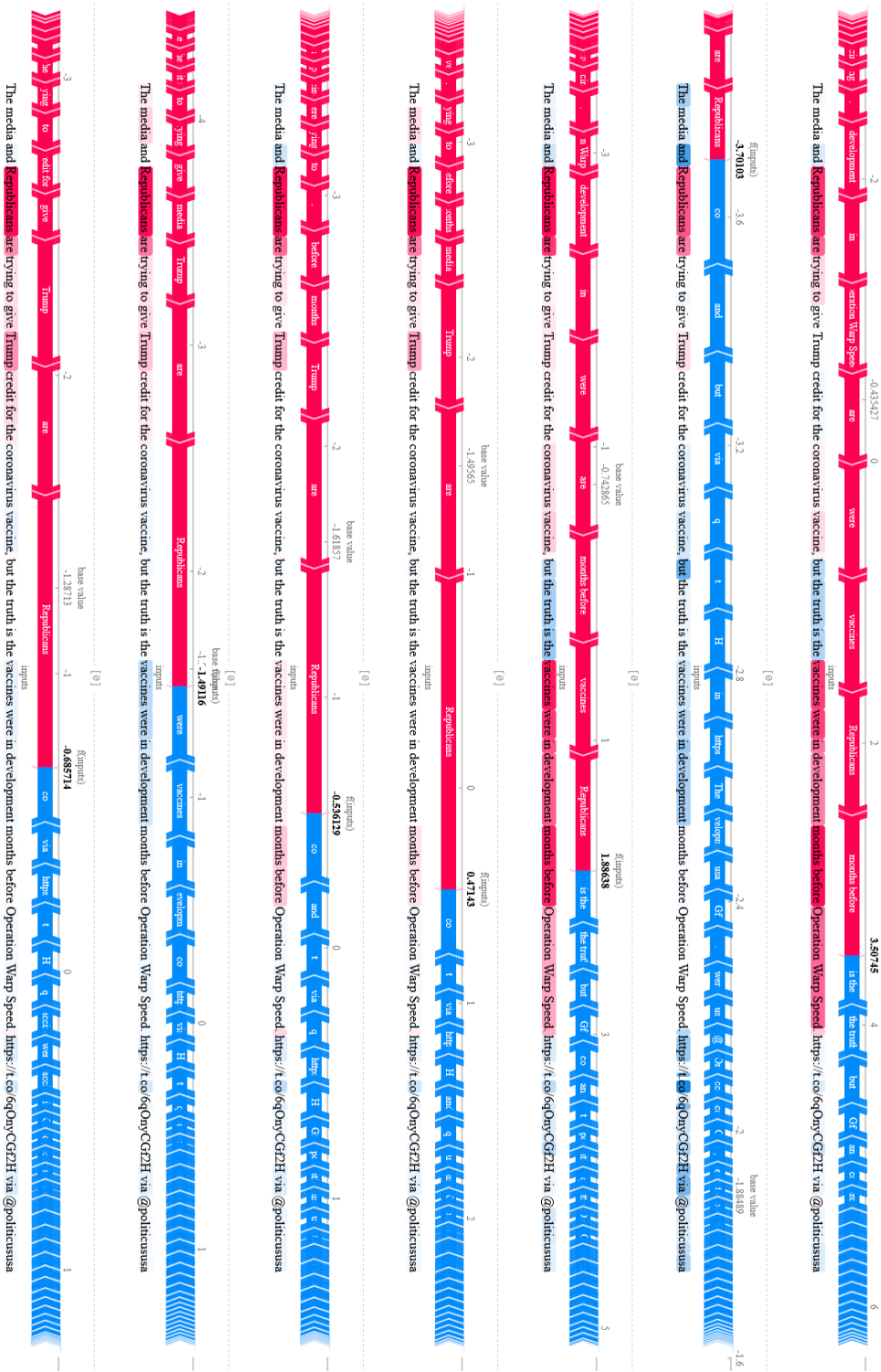


Figure B.7: Tweet ID: 927 SHAP Overview

Actual answer: Yes, No, Yes, Yes, No, No, No & Predicted answer: Yes, No, Yes, Yes, No, No, No. Red highlights are positives to the prediction and Blue are negatives to the prediction

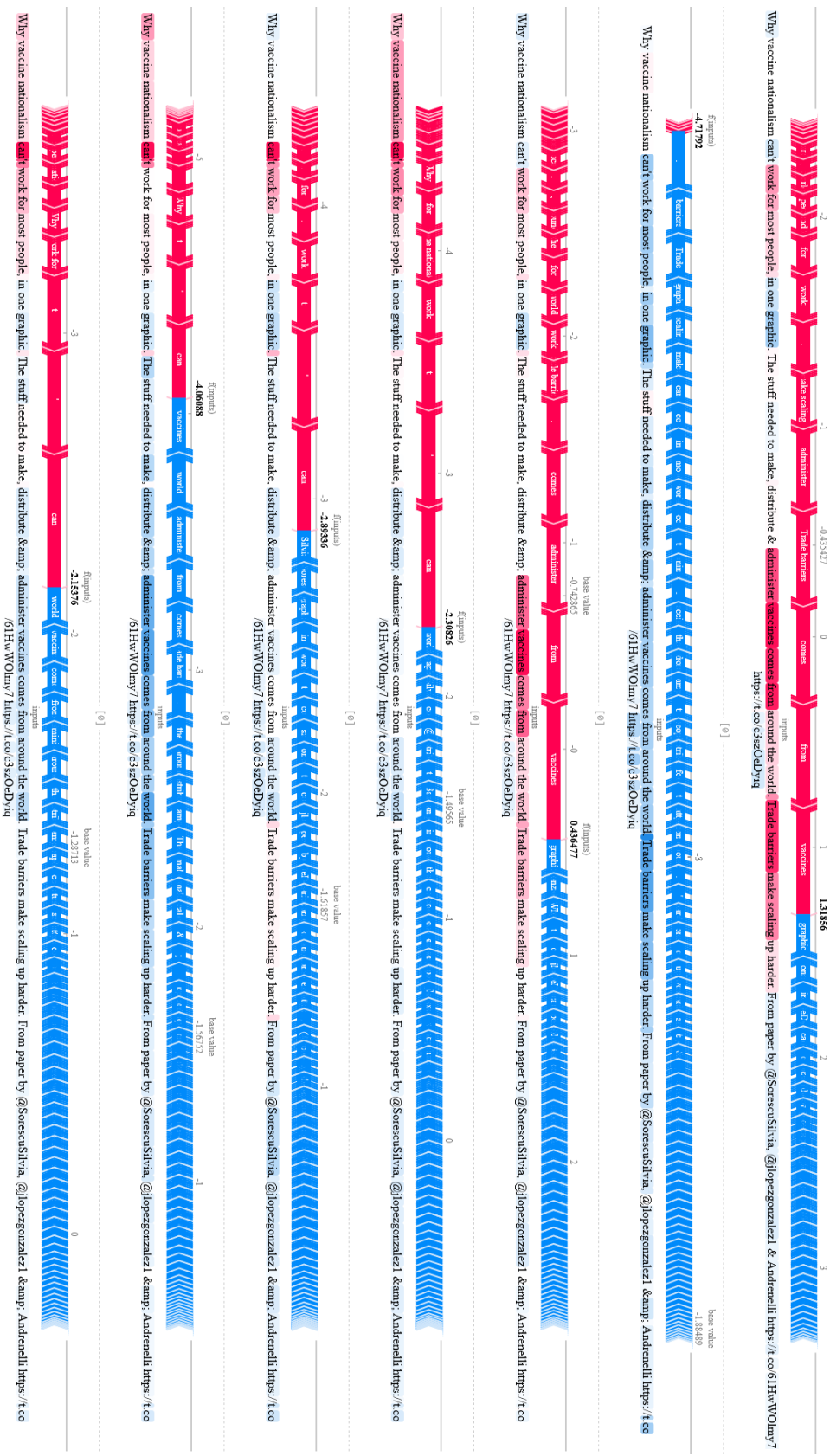


Figure B.8: Tweet ID: 1020 SHAP Overview

Actual answer: No Nan Nan Nan Nan No No & Predicted answer: Yes No Yes No No No No

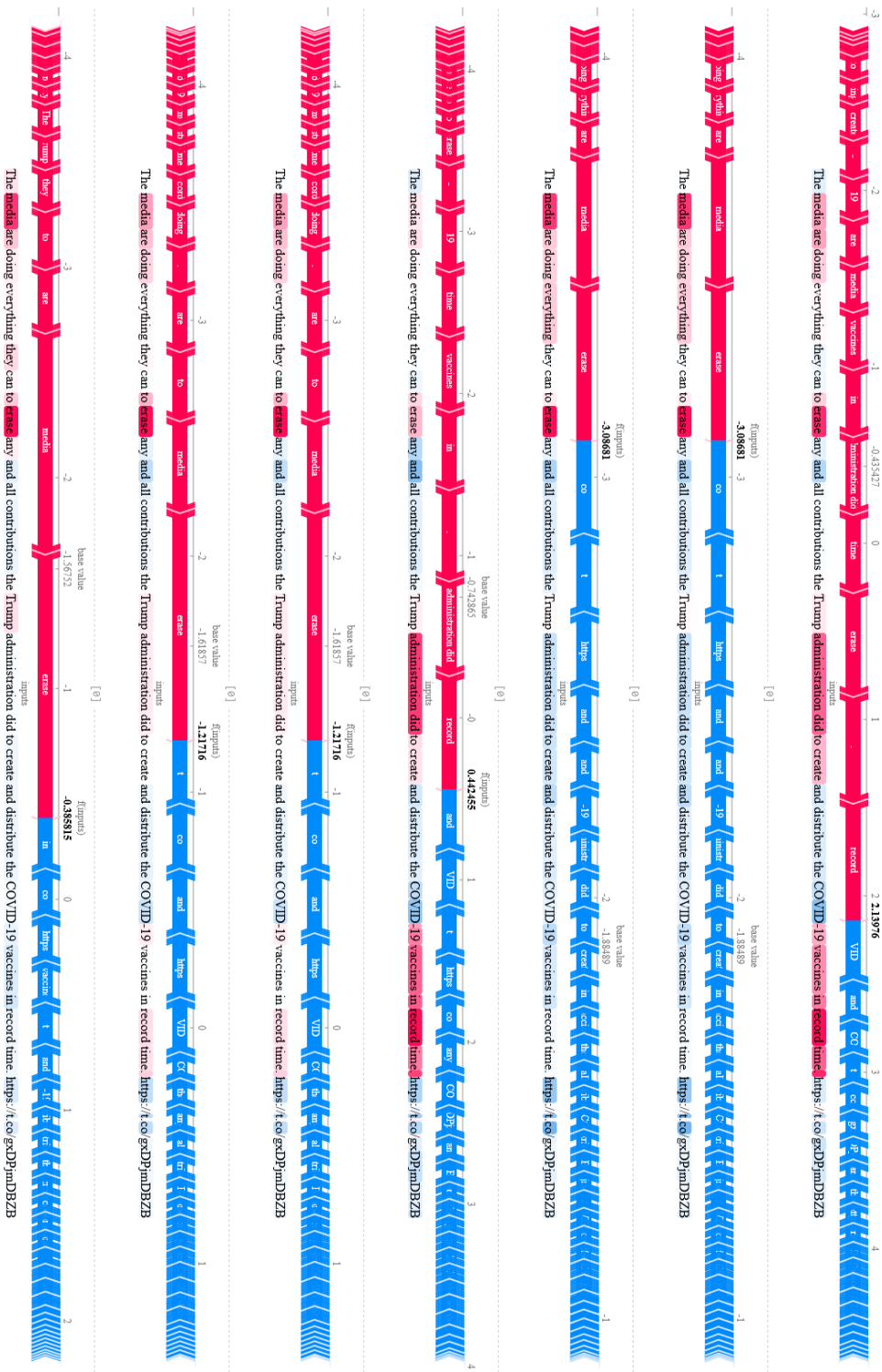


Figure B.9: Tweet ID: 1066 SHAP Overview

Actual answer: Yes Nan Yes Yes Yes Yes Yes & Predicted answer: Yes No Yes Yes No No No

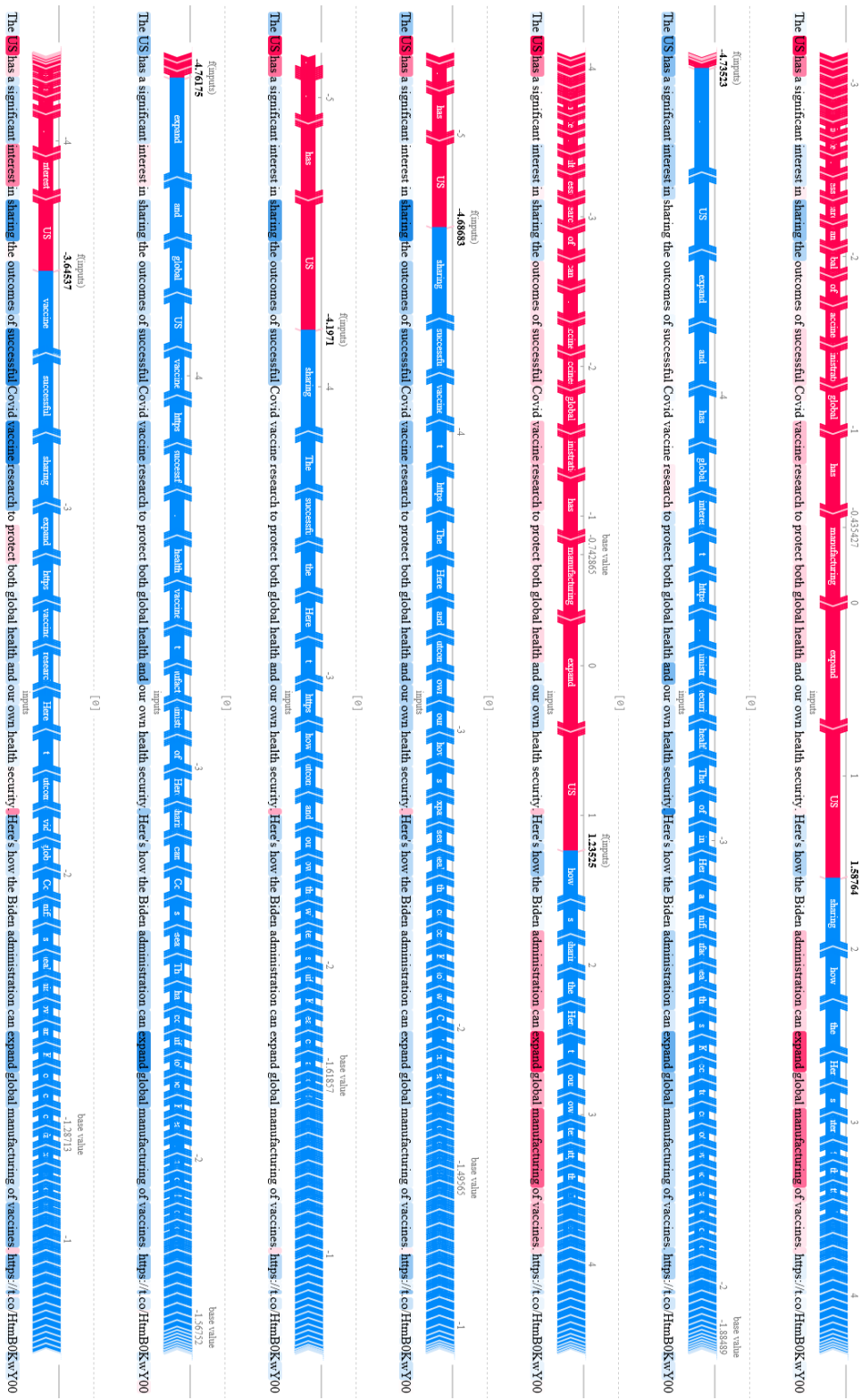


Figure B.10: Tweet ID: 962 SHAP Overview

Actual answer: No Nan Nan Nan Nan No No & Predicted answer: Yes No Yes No No No No