AN AUTOMATED PRIVACY INFORMATION DETECTION APPROACH FOR ONLINE SOCIAL MEDIA

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

Supervisors

A/Prof. Quan Bai

Dr. Weihua Li

July 2019

By

Jiaqi Wu

School of Engineering, Computer and Mathematical Sciences

Copyright

Declaration

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 candidate

Acknowledgements

The thesis research of my Master degree is totally a different challenge to me. It differs from previous teaching courses because its requirements of the abilities of critical thinking and problem-solving are much stricter. Moreover, research is also an iterative procedure, so patience and persistence are also essential when it comes with persistent failure.

I think I cannot complete it with much guidance and help from lots of people.

First and foremost, I want to express thanks to my primary supervisor A/Prof. Quan Bai. He keeps assistance with my research all the time with a high degree of enthusiasm, motivation, and patience. Through the procedure of literature review, he inspires me to broaden my mind and guides me to find my own research direction. Moreover, his immense knowledge really does a huge help to me in the steps of methodology and experiment. When I face some problems, he always gives me significant encouragement and advice by taking his own experience as an example.

Secondly, my second supervisor, Dr Weihua Li also supports me a lot continuously. His technical experience in this field is also of great assistance to my research. He also provides me with some support and tips during the writing of the thesis. Moreover, when I feel stressful during the procedure, he is keen to encourage me a lot to cheer me up.

Thirdly, I want to thank my parents for their financial support of my study in New Zealand. Their unconditional love and encouragement make it possible for me to focus

on my research without financial pressure.

Finally, my gratitude also goes to all the members of the Department of Computer Sciences in Auckland University of Technology for their provided facilities and support, which support this thesis and make it successful.

Abstract

Online Social Networks (OSNs) have become ubiquitous in the activities of people recently. However, a large number of disclosing private information are posted by online social network users unconsciously every day, and some users may face undesirable consequences, e.g., identity theft. Consequently, the significance of privacy information detection for users of OSNs turns out to be important. A large number of studies have been dedicated to corporate privacy leakage analysis. Whereas, there are very few studies that detect privacy revealing for individual OSNs users.

With this motivation, this thesis aims to propose an automated privacy information detection approach to effectively detect and classify privacy revealing information for individual users. It comprises two steps: detecting privacy information leaks and classifying them into fine-grained categories. In the first step, a deep-learning based model is built to recognise privacy-related entities in a real-world data set, which has achieved a considerable performance based on the experimental results and case studies. In the second step, a semantic phrase similarity degree approach is developed to automatically classify privacy-related entities into fine-grained privacy entities based on a built privacy domain ontology. Finally, extensive experiments are conducted to validate the proposed privacy information approach, and the empirical results demonstrate its superiority in assisting OSNs's users to avoid the privacy leakage.

This work provided a complete approach to handle privacy information detection on online social networks, which is essential for individuals to mitigate their privacy

leakage.

Publications

Weihua Li, Jiaqi Wu and Quan Bai. (2019): An Automated Privacy Information Detection Approach For Protecting Individual Online Social Network Users. In Proceedings of the 33rd Annual Conference of Japanese Society for Artificial Intelligence (JSAI), Japanese Society for Artificial Intelligence (JSAI) (Accepted)

Jiaqi Wu, Weihua Li, and Quan Bai. (2019): Privacy Information Classification: A Hybrid Approach, In Proceedings of the 4th International Workshop on Smart Simulation and Modelling for Complex Systems (SSMCS 2019) (Accepted)

Contents

List of Tables

List of Figures

Chapter 1

Introduction

According to Seerden, Salmela and Rutkowski (2018), Internet users have accounted for half of the world's population and 42 % of them are active Online Social Networks (OSNs) users in 2018. With these numbers growing rapidly, OSNs have become of one of the essential channels for social interactions and communications (Seerden et al., 2018). However, it is easy for cybercrimes to be perpetrated on online social media platforms because OSNs allow people to interact with a large number of anonymous users with various backgrounds. While the cybercrimes caused by online social media have kept increasing, there have only been limited studies focusing on detecting privacy information to protect individual OSNs users. The purpose of this thesis is to address this by proposing an automatic privacy information detection approach to protect individual OSNs users.

In this Chapter, Section 1.1 is structured to introduce the background of the thesis, including Web 2.0 (Subsection 1.1.1), Online Social Networks (Subsection 1.1.2), and privacy issues in OSNs (Subsection 1.1.3). Then the motivations and objectives of this thesis are presented in Section 1.2. It comprises two parts, i.e., the motivations of a deep learning based approach (Subsection 2.2.1) and the ontology-based classification approach (Subsection 2.2.2). In Section 1.3, the research methodology of the thesis is

described. Then the main contributions of the thesis are outlined in Section 1.4. Lastly, the organization of the thesis is structured in Section 1.5, which aims to present briefly the contents flow of this thesis.

1.1 Background

1.1.1 Web 2.0

Since the beginning of the 21st century, users of the Web are able to create content instead of just viewing content on the Internet because of the advances and development of technologies (Cormode & Krishnamurthy, 2008). This movement is named as "Web 2.0" or "Social Web". A Web 2.0 website allows users to interact and communicate as the creator of user-generated content through electronic devices, e.g., laptops or smartphones. This differs from the Web 1.0 websites where people were only able to view and find content on the Internet (Cormode & Krishnamurthy, 2008).

1.1.2 Online Social Networks

In the era of Web 2.0, the biggest change is the emerging existence and popularity of OSNs, e.g., Twitter, Facebook, and Instagram. The emergence of OSNs has transformed the way people communicate and share information with each other in their daily lives. Before the era of OSNs, people mostly interact with people they know personally because of the limited communication ways. Currently, users can write blogs, post photos and comment on the posts generated by other users on Online Social Media platforms (Ge, Peng & Chen, 2014). Because of this, they communicate and share information with a large audience and sometimes that numbers are much more than they intend.

Advanced technology in portable electronic devices is also a significant factor

contributing to OSNs becoming ubiquitous in the activities of people. As is mentioned before, OSNs were originally a platform that people could only access via a desktop or laptop. However, with the advent of smartphones, online social media released mobile app versions and developed them into standalone mobile apps. Consequently, it is more convenient for users to access OSNs to write posts or edit profiles (Aldhafferi, Watson & Sajeev, 2013). Moreover, the more accessible OSNs are, the more information users will share online, resulting in its constant presence in daily lives (Coyle & Vaughn, 2008).

The increasing number of people joining them is another factor causing the ubiquitous characteristic of OSNs. For example, Facebook has rapidly grown to become one of the world's most popular OSNs since it was launched in 2004 as an online social media platform for students at Harvard University. According to Batra, Sidhu and Sharma (2018), in 2013, Facebook already had more than a billion active users worldwide. Kemp (2018) indicates that half of the population on the world was an active Internet user in 2018, and over 42 % of them have become active users on online social media. Moreover, the number of OSNs users maintains its rapid speed to increase.

However, the sudden popularity of OSNs, their constant presence in people's daily lives, and their ability to connect a large number of people around the sharing of potentially sensitive information have made OSNs rich sources of personal information and this gives rise to a number of privacy issues.

1.1.3 Privacy Issues in Online Social Networks

OSNs also bring serious privacy issues. Users don't pay much attention to the massive amount of private information, which can be accessed publicly through OSNs, e.g., likes and dislikes, email addresses, education background, hometown, activities attended and anything else. For users, sharing this information allows them to maintain long-term

relationships with friends or to communicate with people with similar interests.

However, users may expose themselves to a wide range of observers, which include not only relatives and close friends, but also strangers and even stalkers. In other words, the readers of their posts are anonymous. This will raise serious cybersecurity issues if their private information is abused.

There are several main negative consequences caused by privacy information leakage on OSNs as follows.

1. Cyberstalking: Cyberstalking has been increasingly prevalent. LeFebvre, Blackburn and Brody (2015) indicate that a large number of users conduct online surveillance in a romantic relationship. More seriously, some people can gain control of their target users by gathering their personal information leaked on OSNs (McFarlane & Bocij, 2003), which will cause harm to the target users.

2. Identify theft: The quantities of personal information can be collected for identify theft, which may cause huge financial loss to users (Humphreys, Gill & Krishnamurthy, 2010). A report conducted by the Department of Justice in the USA shows that 17 million cases of identity theft caused almost 25 billion dollars losses in 2015 (Francia III, Hutchinson & Francia, 2015).

3. Phishing: Cybercriminals can easily conduct phishing attacks by collecting personal revealing information of the targeted victim from OSNs (Dewan, Kashyap & Kumaraguru, 2014). Then the attackers aim to craft a scenario that looks realistic after collecting sufficient information. Because of the scenario, it is very simple for the attackers to obtain the trust of the victim. Then they use the trust gained to trick victims into surrendering privacy information, e.g., credit card details.

4. Romance fraud: Romance fraud crimes are another type of cybercrimes facilitated by privacy leakage on online social media, which is also called "catfish" (Derzakarian, 2017). It refers to that victims are lured into a romantic relationship by being attracted from some fabricated personas. Those fake personas are built by cybercriminals through collecting leaked personal information such as likes and dislikes, and then victims may face serious financial loss or psychological damage.

In summary, privacy-related information leaking on online social media platforms can be used by malicious users and exploited by cyber crimes such as phishing, identity theft, romance fraud, and cyberstalking (Seerden et al., 2018).

1.2 Research Motivations and Objectives

Consequently, we urgently need to find an effective and practical solution to detect privacy leaks on OSNs in order to reduce the caused negative consequences, especially serious crimes. It is also necessary to have a tool to assist general users to make better use of OSNs and protect them from leaking privacy information (Wang et al., 2011). Hence, it is essential to detect privacy leakage in OSNs and remind individual online social network users before posting any privacy-related messages to the public (Hasan, Habegger, Brunie, Bennani & Damiani, 2013).

Under this motivation, in this thesis, we propose a hybrid privacy detection approach for individual users of OSNs. It is composed of two parts: developing a deep learning-based approach to detect the privacy information on online social data and classifying them into fine-grained categories based on a privacy domain ontology.

1.2.1 Deep Learning Based Detection

The first part of the hybrid privacy detection approach is based on a deep learning model. The research motivations of using the deep-learning based detection approach are as follows:

1. To capture privacy-related information from posts on OSNs.

2. To remind users of the possibility of privacy leakage.

To achieve the motivations mentioned above, we exploit a deep learning model in this thesis. Deep learning models can build the connection between raw data and the outcomes directly as an end-to-end learning approach which can be used in extracting privacy-related information (LeCun, Bengio & Hinton, 2015). Scanning through some sample posts on OSNs, the individual privacy information and the privacy rules can be formally defined. Based on the privacy definition, a deep learning-based approach can be developed and utilized to recognize privacy-related entities for individual online social media users. Then the users can be reminded of the possibility of privacy leakage, based on the defined privacy rules.

1.2.2 Ontology Based Classification

However, two major limitations exist in the deep learning based detection approach. Firstly, it can only remind users of the possibility of privacy leakage. As a result, detailed leaking information in terms of what kind of leakage is ignored. Secondly, the privacy model is not conceptually modelled or presented. To solve the two problems, the research motivations of utilising the ontology-based classification approach are as follows:

1. To organise the privacy concepts in the form of taxonomy or hierarchy.

2. To remind individual users of what is to be disclosed instead of simply giving general information.

Based on the motivations above, we present an ontology-based classification approach to classify fine-grained privacy information further. Ontology models can conceptually reflect the domain-specific knowledge in the form of terms. Moreover, an ontology demonstrates two major advantages, i.e., shareability and reusability (Zhao, Dong & Peng, 2009). Therefore, ontology-based privacy models can be easily extended and applied to various OSNs (Mitra, Liu & Pan, 2005). Moreover, as ontology organises the concepts in the shape of taxonomy or hierarchy based on a pre-defined natural relationship, it is suitable to introduce ontology into the research of privacy-related information extraction and classification.

By scanning all the privacy-related entities recognised by the deep learning model, a privacy domain ontology can be formed which includes extracted representative terms based on privacy concepts. Then the ontology-based approach is proposed to calculate the semantic similarity degree between the entities extracted from the deep learning model and the representative terms. Through this approach, the specific categories of privacy disclosure information can be advised to individual online social media users.

1.3 Research Methodology

The research methodology of the thesis is summarised and presented in Fig. 1.1. Firstly, we review the existing researches regarding privacy leakage problems and detection approaches on OSNs. This step aims to investigate the research gap, i.e., the research problems addressed in this thesis. Then the objective of the thesis is demonstrated, i.e., to detect privacy leakage on OSNs and remind online social networks users in detail before posting any privacy-related messages. Secondly, the individual privacy information on OSNs has been formally defined according to the objective of the

thesis. In the third step, we establish a hybrid privacy information detection approach containing a deep learning model and an ontology-based approach. In the fourth step, the experiments are conducted by collecting real-world dataset, pre-processing data, and processing data. The experimental results are discussed to validate the proposed approach in the fifth step. Based on the results of the experiments, we can adjust the descriptions and definitions of the problem to optimise our model to detect privacy leaking information. Finally, the model can be built to solve the problems addressed by this thesis.

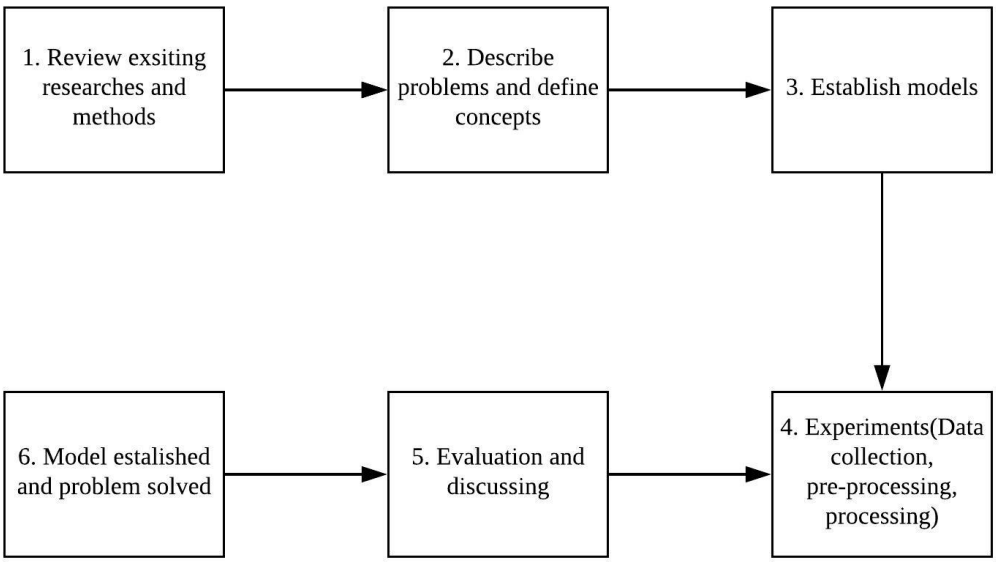

Figure 1.1: Research Methodology

1.4 Contributions of The Thesis

For privacy information detection, most researchers investigate the problem from the perspective of corporations; whereas few studies pay attention to privacy information detection for individual OSNs users. Therefore, in this thesis, we make three major contributions.

1. Definition of privacy information on OSNs: The first contribution is that we define the privacy information that users tend to publish on OSNs. It is defined as a sequence of words, stating or implying any personal information about an individual user, preferences, events that he or she involved. Then we define privacy rules to judge whether the message is private, based on a word sequence.

2. An automated privacy detection approach for individual OSNs users: Secondly, we propose an automated privacy detection approach for protecting individual OSNs users based on the definition. In the entire approach, the deep-learning based model to recognise privacy-related entities plays a significant role; this is the second contribution of the thesis.

3. A privacy domain ontology building: In the entire detection approach, a privacy domain ontology has been built to classify fine-grained privacy information and this is the third contribution of the thesis.

1.5 Thesis Organisation

The rest of the thesis includes the following chapters.

Chapter 2 reviews the existing research work regarding data leaks on OSNs from the perspective of corporations and individuals respectively and then indicates the research gap. It also evaluates the related studies regarding two key parts of the privacy information approach we proposed, including Named Entity Recognition (NER), the rationale of the state-of-art model for NER, and some studies about ontology-based classification approaches.

Chapter 3 introduces the hybrid privacy detection approach for individual OSNs users in detail. In this approach, there are two key parts: *deep learning-based detection approach* and *ontology based classification approach*. Firstly, the rationale for utilising a

hybrid model of both deep learning and ontology-based classification models is clarified. Secondly, the deep learning based privacy detection model is presented and evaluated in two experiments. The result of the first experiment is based on the evaluation metric (precision, recall, and F1 value). Tuning of parameters and word embedding approach is adjusted in the first experiment to get the best model performance. In the second experiment, three case studies are selected to evaluate the effectiveness of the proposed detection model.

In Chapter 4, the ontology-based classification approach is presented to classify privacy information into fine-grained categories, which can remind OSNs' users in detail. Firstly, privacy-related representative terms are extracted from the results of the deep learning model in Chapter 3. Secondly, a privacy domain ontology can be built according to the representative terms. Thirdly, a semantic similarity degree approach is proposed to do fine-grained classification further. Then the result of experiments demonstrates that the proposed hybrid privacy information detection approach in this thesis shows a considerable performance because of the high accuracy in each type. And lastly, the distribution of different categories of privacy revealing information is presented and discussed.

The final chapter, Chapter 5, concludes the findings of this thesis, as well as the limitations and potential directions related to this research for future work.

Chapter 2

Literature Review

2.1 Introduction

OSNs are used as a tool to facilitate crimes such as identity theft, romance scams, and cyberstalking because of the existing privacy leaking problems. Consequently, privacy revealing detection in OSNs has been a hot topic in recent years. There is a large quantity of research focusing on the topic. Some existing research papers related to privacy information detection are reviewed and discussed in Section 2.2 of this chapter.

We categorise them into two types in terms of the objectives they attempt to address the privacy detection problem for: (1) for corporations, especially for the health care industry (Subsection 2.2.1); (2) for individuals, including Twitter users and other OSNs platforms users (Subsection 2.2.2). We review and analyse these privacy detection researches on OSNs from several aspects: the objectives they attempt to achieve, the detection approaches they propose, and the conclusions they obtain. In addition, we also discuss what types of privacy leakage they are trying to detect. Finally, we point out the research gaps existing in the area of privacy information detection on OSNs, which is the objective of this thesis.

In Section 2.3, we review studies about Named Entity Recognition(NER), which

is the main technique considered the proposed approach of the thesis. According to related studies, we categorize NER approaches into three kinds of types: (1) Traditional approaches which are based on rules and dictionaries (Subsection 2.3.1); (2) Machine learning based approaches (Subsection 2.3.2); (3) Deep learning based approaches (Subsection 2.3.3). We compared their advantages and disadvantages and chose the deep learning based approach to best achieve the objective of this thesis.

In Section 2.4, the rationale of the state-of-art model in the deep learning based NER model, i.e., the Bi-LSTM CRF model, is demonstrated. We describe the rationale from the Bi-LSTM model (Subsection 2.4.1) and the CRF layer (Subsection 2.4.2). Then we illustrate the reason why we combine them together to recognise the privacy-related entities in this thesis. Lastly, in Section 2.5, some studies related to ontology-based classification approaches in other fields are analysed. Then we explain why it is essential to build a privacy information ontology in this thesis in order to further classify privacy-related information into fine-grained categories.

2.2 Existing Privacy Detection Approaches on OSNs

Privacy revealing detection in OSNs has attracted a large amount of research. However, most of it focuses on how private or public corporations can maintain adherence to policies and regulations that promote privacy protection in accordance with their industry.

2.2.1 Privacy Detection Approaches for Corporations

Among the studies about privacy detection approaches for corporations, most of them focus on the health care industry because of its heavy regulation.

Naslund, Grande, Aschbrenner and Elwyn (2014) indicate individual users with Severe Mental Illness (SMI) tend to use YouTube to share their mental health care

advice by uploading videos and connecting with other users actively. On one hand, this kind of action may improve psychosocial outcomes, including learning from other individuals and minimising a sense of isolation. On the other hand, it may cause disclosure of personal health information, which may cause negative comments about users' videos and adversely influence the mental states. They also recommend that OSNs like YouTube should take action in compliance with relevant privacy regulations.

According to Li (2014), Health Social Network System (HSNS), concern over the privacy of published health care data has been caused. They point out that systems rather than techniques are the key to protecting health-related private data through conducting a real-world dataset case study. They also indicate that HSNS corporations are required to meet some key system requirements to protect the health care data of OSNs users.

To specifically protect the privacy information of online patients and health care plan participants, the Health Information Technology for Economic and Clinical Health (HITECH) Act has been enacted (Pathman, Crouse, Padilla, Horvath & Nguyen, 2009). Similarly, the Children's Online Privacy Protection Act (COPPA) aims to protect the personal information of online social media users who are children under the age of 13 (Francia III et al., 2015). In addition, there are other International privacy laws to ensure that corporations must provide adequate protection of private data about OSNs users. The best-known measure is the General Data Protection Regulation (GDPR).

According to Seerden et al. (2018), the European Union(EU) has promulgated GDPR for organisations to handle privacy information for privacy-related data protection. They conduct multiple case studies to analyse what actions organisations can take to prevent privacy violations in order to comply with GDPR. Similarly, Henderson and Snyder (1999) propose a protection framework that organisations can adjust to detect potential privacy leak problems by analysing privacy policy and self-regulation of those organisations. Through the experimental results, they evaluate the model that

can prevent some negative consequences caused by privacy revealing.

Moreover, Walczuch and Steeghs (2001) conduct an exploratory survey about changes to multinational corporations for detecting privacy leakage due to the processing of personal data in GDPR. Through the survey analysis, they indicate that the new data regulation affects mostly multinational corporations that process personal customer data across the EU borders. Then they draw a conclusion that those companies are facing more problems to comply with the requirements of GDPR.

In summary of privacy detection approaches for corporations, Ohlhorst (2012) shows that an organisation is supposed to keep data security to comply with its own goal. They analyse multiple studies on privacy detection approaches for organisations to determine the privacy protection requirements and the solutions of privacy leak.

From the above studies, we can see that there are a large number of studies focusing on privacy detection approaches from the perspective of organisations, including analysis of government privacy regulations to protect both patients and children. However, there are few studies about privacy detection for individual OSNs users; these will be considered below.

2.2.2 Privacy Detection Approaches for Individuals

However, there is only a limited amount of research focusing on privacy leakage on OSNs users.

Humphreys et al. (2010) indicate people are careful about sharing specific personal information like the home address or their phone numbers, which means those kinds of information are rarely found in tweets. However, users can easily disclose unconsciously where or when they are. For example, a tweet saying that user's family will go on holiday implies no one stays at home; this may result in a burglary. Therefore, the privacy information that OSN users provide should be queried and confirmed before

sharing with the public.

Gomez-Hidalgo et al. (2010) proposes a machine learning classification approach by adopting Named Entity Recognition technology. It can detect entities of some categories like electronic devices and brands in tweets and users will be prevented to post these privacy-related entities. However, this study considers no means detection of privacy information about a user themselves.

Mao, Shuai and Kapadia (2011) present a privacy classifier for three kinds of sensitive tweets, i.e., drunk, holiday and disease tweets. Through the classifier, they classify tweets into binary results successfully, i.e., sensitive tweets or nonsensitive tweets. Moreover, they investigate who reveals privacy information and they do cross-cultural analysis by comparing the privacy revealing distribution in the United Kingdom, the United States of America and Singapore. The limitation of their work is that they just consider the three types and they extract the topic tweets in advance, which means they have known the privacy types of tweets. Whereas, it is impossible for individual OSNs users to get a clear understanding of the type of their privacy leakage.

There are also some studies that investigate privacy issues in other online social networks in general.

Bhagat, Cormode, Srivastava and Krishnamurthy (2010) investigate the privacy-preserving problem in evolving networks. They propose an approach to anonymize a dynamic network that the privacy of users is preserved when adding new nodes and edges to the published network. Moreover, the evolution model is implemented using a link prediction algorithm. Finally, they obtain the conclusion that privacy information loss can be eliminated in the predicted social graph.

Acquisti and Gross (2006) investigate the privacy concerns of users on Facebook. They take a group of Facebook users as a representative sample and compare the survey results with the information extracted from the Internet. From the results of the experiment, they indicate that Facebook users reveal a large amount of privacy

information and most of them think they can control privacy leakage by restricting external access to Facebook.

Dwyer, Hiltz and Passerini (2007) make a case study on trust and privacy issues between MySpace and Facebook. They conducted a survey about perceptions of trust and privacy concerns on OSNs. Then they analysed them comparing the general use of the two sites. Through the correlation analysis, they find the levels of privacy concerns on the two sites are similar.

From the researches studied above, we can see even though there are several papers about analysing privacy revealing on OSNs, very few studies investigate how to detect individual privacy information and protect individual OSNs users from online privacy leaks. Secondly, among researches, most contemporary privacy information classification approaches aim to detect some specific categories of privacy information on OSNs or perform a binary classification, i.e., sensitive or non-sensitive, rather than identifying and classifying privacy information for the end users.

Therefore, in this thesis, instead of assisting the organisations, we target the individual online users and keep them away from privacy leakage. As almost all the posts by users are unstructured data, the information extraction method plays a pivotal role in the proposed framework.

2.3 Named Entity Recognition

As previously indicated, privacy information extraction is essential in the proposed detection approach. Moreover, Named Entity Recognition (NER) is an important method for extracting domain-specific information from texts (Nadeau & Sekine, 2007). It is intended to identify words or phrases as pre-defined tags that describe domain-specific concepts of interest. For example, Gomez-Hidalgo et al. (2010) proposed an effective mechanism that is capable of detecting some types of named entities, e.g.,

company, brand of electronic devices, and person, using the NER technique. Given the context of the privacy detection domain, NER can assist users in identifying privacy-related entities after being given sufficient training.

2.3.1 Traditional Approaches

There exist various kinds of NER methods. Traditionally, dictionary-based (Aronson, 2001) and rule-based approaches (Farmakiotou et al., 2000) can be used to identify domain entities. However, the dictionary-based approach matches the phrases with synonyms existing in the dictionaries, which requires frequent updating of the dictionaries with new concepts and synonyms. Moreover, they only identify the existing entities in dictionaries, which is another limitation. Similarly, the rule-based approach has its own limitation, i.e., it needs to construct rules manually, in which it is time-consuming.

2.3.2 Machine Learning Based Approaches

To be less dependent on dictionaries and manually created rules, machine learning based approaches have been popular in NER recently. Machine learning begins from the early research in Artificial Intelligence, which gives the computer the ability to learn without specific programming (LeCun et al., 2015). A large amount of machine learning based algorithms emerged these years, e.g., decision-tree algorithms and clustering algorithms.

For NER, machine learning approaches build classifiers using relevant algorithms for each class of entity by training large texts with its gold standard long-form labelled by professionals (Lakshmi, Panicker & Meera, 2016). In Fig. 2.1, a classifier is a significant part of a machine-learning pipeline and is utilised for identifying new data categories. It analyses the marked training data set based on known information about the class categories. According to Fig. 2.1, the feature extractor is used to convert raw

data into representation types that can be used by machine learning classifiers because classifiers cannot directly identify their patterns using raw data (LeCun et al., 2015). Moreover, feature extractors differ a lot regarding objectives. Consequently, building a variety of feature extractors still requires considerable work and expertise. (LeCun et al., 2015).

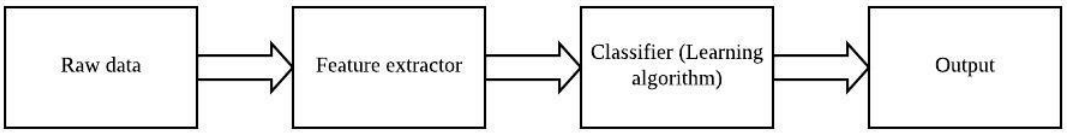

Figure 2.1: The Workflow of Machine Learning

In the procedure of machine learning training in NER, some text-based features are utilised by researchers like part-of-speech (POS) and position feature. Lakshmi et al. (2016) train a Support Vector Machine (SVM) classifier for each class because the SVM algorithm generally has excellent performance on multi-dimensional data. They evaluate their experiments on the data sets and achieve the result with an 83.82 % accuracy value. Moreover, word embedding technology is proved as a dominant feature to improve the performance in NER.

2.3.3 Deep Learning Based Approaches

However, deep learning, which is a branch of machine learning algorithms, does not need the feature extractors (LeCun et al., 2015). "Deep Learning" is actually a series of new structures and methods that are developed to allow a neural network with a large number of layers to train and work (LeCun et al., 2015). The neural network can be divided into three parts, namely the input layer, the hidden layers and the output layer. The input can be received from the input layer of the neural network, and be processed through multiple hidden layers, and the result is generated at the output layer.

In general, the network structure of deep learning is also a multi-layer neural network.

Therefore, the deep learning model can automatically learn the characteristics from the labeled training data set through the neural network. In other words, it can directly establish the connection between the raw data and the outcomes during the training process in Fig. 2.2, which is much easier than machine learning approaches.

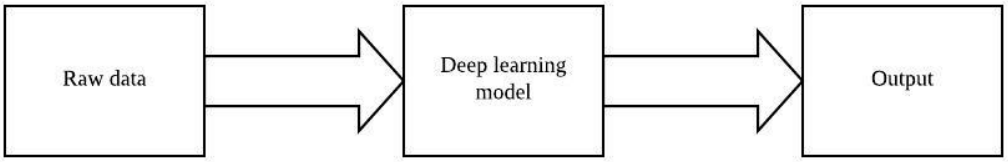

Figure 2.2: Deep Learning Workflow

2.4 Bi-LSTM CRF Model

Among studies of deep learning based NER approaches, Ren, Teng, Li, Chen and Ji (2017) propose a Recurrent Neural Network (RNN) framework based on Long-Short Term Memory (LSTM) nodes. Their model proves that the LSTM model outperforms than classical machine learning models including SVM in NER. Nowadays, the Bi-LSTM with Conditional Random Field (Bi-LSTM CRF) model becomes more popular as it achieves more promising results than LSTM model (Lample, Ballesteros, Subramanian, Kawakami & Dyer, 2016). Therefore, we utilise the Bi-LSTM CRF model for privacy-related entities extraction.

2.4.1 LSTM and Bi-LSTM

In the deep learning based approach, the traditional neural network model just connects from the input layer to the multiple hidden layers, and then to the output layer. There is no connection between the nodes in the same layer, and the propagation of the

network is sequential(Lample et al., 2016). However, this structure may be not suitable for Natural Language Processing (NLP). If a task needs to predict the next word, the previous words of the word will need to be known, because of the semantic meaning in a sentence (Collobert & Weston, 2008).

RNN is demonstrated to extract named entities, which can solve the problem (Yepes, 2017). It has an advantage that the nodes in hidden layers can store historical memory because of its folded structure. The nodes of the hidden layers are no longer independent of each other, and the output is based on all previous calculations.

However, the long-term dependency learning problem still exists in RNN (Wen et al., 2015). Sometimes it may be difficult to transfer information from an earlier time state to a later time state because of the length of the sequence. Therefore, RNN may miss important information if a sequence requires to be processed, and the result will be biased towards its most recent inputs in a sequence (Lample et al., 2016).

LSTM can solve the long-term dependency learning problem because the structure of the nodes in the hidden layers are much more complicated than nodes in hidden layers of RNN, which can be called 'cell' in Fig. 2.3 (Huang, Xu & Yu, 2015). The complicated structure of cell makes the information can be reserved.

In Fig 2.3, according to Huang et al. (2015), U represents the weight matrices from the input layer to the hidden layer and V represents the weight matrices from the hidden layer to the output layer. The cell in the hidden layers includes a gating mechanism, in which three gates can be controlled whether to keep the memory information (Huang et al., 2015). The gates are named as input, forget and output gates, which can be represented by i, f, and o respectively in the structure of the cell A. W and b are utilized to represent the weight matrices and bias vector respectively.

The forget gate is utilised to decide whether to forget some information from the former state C_{t-1}, which decides what information should be discarded or retained (Huang et al., 2015). It utilizes an activation function, i.e., σ function, to calculate

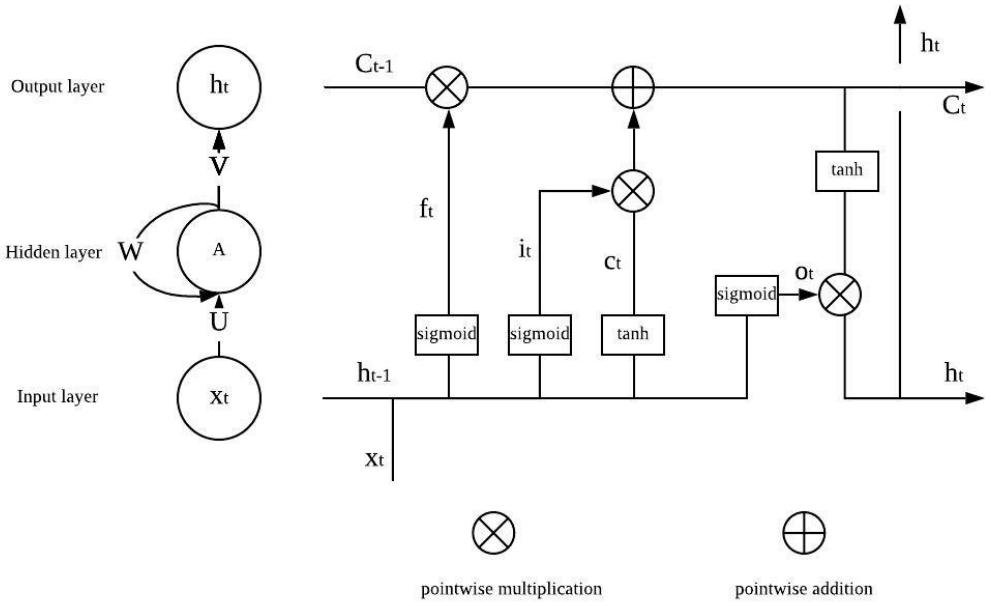

Figure 2.3: LSTM Cell

former state's output h_{t-1} and current input value x_t to get the output value f_t (Huang et al., 2015). Then the calculation of the forget gate's output value can be shown:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{2.1}$$

The input gate is used to decide what information can be saved to the current state C_t considering the current input information x_t, which can prohibit unimportant information from being updated (Huang et al., 2015). First, the former state's output h_{t-1} and the current input x_t are calculated by the σ function (2.2). Secondly, these two values are also calculated by the $\tanh$ function to create a new candidate value, i.e., $\tilde{C}_t$ (2.3). Finally, the updated values of the input gate can be shown below, where the character $\odot$ means elementwise multiplication (2.4) (Huang et al., 2015):

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{2.2}$$

$$\tilde{C}_t = tanh(W_c[h_{t-1}, x_t] + b_c) \tag{2.3}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \tag{2.4}$$

According to Huang et al. (2015), the output gate is utilised to decide how much information can be outputted, i.e., h_t, regarding the combination of C_t, h_{t-1} and x_t, which can be calculated as depicted:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{2.5}$$

$$h_t = o_t \odot tanh(C_t) \tag{2.6}$$

The bidirectional LSTM (Bi-LSTM) model includes a forward and backward LSTM, which can be utilized to process the sequence from two directions respectively (Lafferty, McCallum & Pereira, 2001). It has the left and right contextual vector for each word, which provides meaningful features for sequences. So the past and future information can be captured. Graves and Schmidhuber (2005) indicate Bi-LSTM model performs better than LSTM model in NER because bidirectional networks are significantly more effective than unidirectional ones.

2.4.2 CRF Layer

Conditional Random Field (CRF) classifier has been employed for NER due to its robustness and reliability (Lafferty et al., 2001). As a conditional probability distribution model, CRF classifier outputs a sequence regarding the inputted sequence by making global optimal predictions.

In CRF, the observation sequence and the label sequence, which can be represented

by x and y respectively (Konkol & Konopík, 2013). Consequently, sequence labelling can be computed as:

$$p(y|x,\lambda,\mu) = \frac{1}{Z(x)} \exp\left(\sum_j \lambda_j t_j\left(y_{i-1}, y_i, x, i\right) + \sum_k \mu_k s_k\left(y_i, x, i\right)\right) \tag{2.7}$$

where the transition feature function in the observation sequence from position $i-1$ to position i can be represented by $t_j\left(y_{i-1}, y_i, x, i\right)$. $s_k\left(y_i, x, i\right)$ represents a state feature function when the label is at position i and the observation sequence followed. In the training dataset, λ_j and μ_k are two parameters to be estimated (Konkol & Konopík, 2013).

$Z(x)$ is a normalized factor as shown:

$$Z(x) = \sum_y \exp\left(\sum_i \sum_k \lambda_k t_k\left(y_{i-1}, y_i, x\right) + \sum_i \sum_k \mu_k s_k\left(y_i, x\right) \quad\right) \tag{2.8}$$

Finally, the output is calculated as depicted:

$$y* = \arg\max p(y|x) \tag{2.9}$$

As we demonstrated before, the Bi-LSTM model can maintain the sequential input information and assume labels independently. However, it does not consider the contextual conditions to assign the word label (Konkol & Konopík, 2013). Regarding the rationale of CRF, the CRF layer can assist the Bi-LSTM model with the contextual conditions function as the final layer, which can bring benefits when considering the correlations between labels (Konkol & Konopík, 2013). Consequently, the Bi-LSTM CRF model is used widely in the NER task recently.

2.5 Ontology Based Classification Approaches

In the privacy revealing information detection on OSNs, detailed leaking information in terms of what kind of leakage is occurring should be not ignored. The privacy model is also supposed to be conceptually modelled or presented.

In the procedure of literature review, ontology models are utilised in related studies because it can conceptually reflect the domain-specific knowledge in the form of terms. There are two major advantages in a domain ontology, i.e., shareability and reusability (Zhao et al., 2009). Therefore, ontology-based privacy models are easily extended and applied to various OSNs (Mitra et al., 2005). Moreover, since ontology organizes the concepts in the form of taxonomy or hierarchy based on a pre-defined natural relationship, it is suitable to introduce ontology into the research of privacy-related information extraction and classification.

From the research above in Section 2.2, we can see even though there are several studies about classifying privacy revealing information on OSNs, most of them pay attention to specific categories by machine learning approaches. Very few studies classify privacy-related information according to a domain privacy ontology and protect individual OSNs users from online privacy leaks. Consequently, an ontology-based classification approach is comparatively a new research field of privacy information classification on OSNs.

Ontology-based classification approaches show outstanding performance in many fields, e.g., online job offers (ul haq Dar & Dorn, 2018). ul haq Dar and Dorn (2018) propose an ontology-based classification mechanism to automatically classify online job offers into IT offers and non-IT offers respectively. The methodology includes extracting concepts from job offers, calculating the minimum threshold for calculation and then the ontology-based approach can be developed. Through the evaluation, the model shows a more than 90% accuracy in classifying online job offers.

Similarly, Galizia (2006) provide a general ontology about trust requirements on semantic web services, which is named as Web Services Trust-Based Selection Ontology (WTSO). They set trust analysis as a classification problem, which can be solved with WTSO. The proposed ontology can present four benefits: generality, open characteristic, trust-based invocation, and explicitness.

Consequently, it is necessary to build a privacy domain ontology into the OSNs, which can solve the privacy information fine-grained classification problem. Moreover, it can be both general and open. Privacy information can be formally described explicitly in an ontology (A. Kim, Hoffman & Martin, 2002).

Gharib, Giorgini and Mylopoulos (2017) make survey research about privacy ontology. They point out although there are some security ontologies for fulfilling security requirements, these studies focus on security rather than privacy(Souag, Salinesi & Comyn-Wattiau, 2012). They present a novel privacy ontology to identify the key concepts and relations to satisfy privacy requirements. However, it aims to meet the privacy requirements from the perspective of software engineers. Actually, there is little research about privacy information ontology which can apply for OSNs.

Therefore, in this thesis, we aim to build an ontology about privacy information on online social media, which can further classify privacy information into specific categories.

2.6 Summary of Literature Review

Privacy information detection on OSNs is an integrated topic that involved different research areas. Firstly, related studies about privacy detection are reviewed, and few of them focus on protecting individual OSNs users. To fill the research gap, we aim to propose a detection approach to keep individual OSNs users safe from privacy leakage.

Then we investigate some studies related to privacy detection techniques. Firstly, we

review the related NER approaches by comparing their advantages and disadvantages. The deep learning approach is suitable for the objective because it can automatically extract features from the labelled training data. Secondly, the rationale of the Bi-LSTM CRF model is demonstrated because it shows an outstanding performance in NER. Finally, some studies related to ontology-based classification in other fields are discussed, which demonstrates the benefits of building a privacy domain ontology. Combining the lack of privacy domain ontology on OSNs, it is necessary for us to build a privacy information ontology to further classify privacy-related information in this thesis.

Chapter 3

Automated Hybrid Privacy Detection Approach

3.1 Introduction

Privacy information detection is a critical issue in Online Social Networks. More importantly, users should be protected from privacy leaking and be reminded before posting any privacy-related messages to the public. For most posts on OSNs, it is difficult for extracting privacy-related entities because there exists a challenge: there are no specific definitions about privacy on OSNs.

Individual privacy information (Subsection 3.2.1) and privacy rules to identify whether a tweet is private or not (Subsection 3.2.2) are defined formally firstly in Section 3.2. Under this definition of personal privacy information, Section 3.3 presents a hybrid privacy detection approach to effectively detect and prevent privacy leakage for individual users, which comprises both the deep learning based NER model and an ontology-based classification approach.

In Section 3.4, the first part of the hybrid privacy detection approach, the deep learning based NER model is illustrated, which is composed of the annotation approach

and the algorithm. Lastly, in Section 3.5, two experiments have been conducted to evaluate the proposed framework by using a real-world data set collected from Twitter.

3.2 Privacy Definition

In this section, in order to make it easier and more accurate to extract privacy-related entities, we give formal definitions of "individual privacy" based on the generic definition of privacy and the nature of privacy leakage on OSNs.

3.2.1 Definition of Privacy Information

The core term, "Privacy", has a very broad meaning, which generally refers to the people's right to protect their personal matters (Gehrke, Lui & Pass, 2011). In this sense, the privacy information is associated with something personal, as well as the matters of past, present, and future. However, there is no clear definition of privacy on OSNs.

To form a better understanding of privacy, there are two main facets of privacy studies: legal privacy and social privacy (Cai, Lu & Zhang, 2010). Both facets define the usage of privacy in the area of philosophy. Legal privacy aims to understand the loss of privacy, and corresponding laws are enacted from the perspective of protecting individuals; this matches the objective of this thesis.

Given the generic definition, legal privacy, and the objective of privacy protection studies, in this thesis, the privacy information that users tend to publish on OSNs is defined as a sequence of words, stating or implying any individual's personal information, preferences, events that he or she involved, which, if revealed, may cause negative consequences to OSNs users.

Based on the definition given above, the privacy information incorporates four categories of entities, i.e., PERSON, TRAIT, PREF, and EVENT as follows:

1. PERSON refers to any expression that identifies a real person.

2. TRAIT represents the personally identifiable information, e.g., home address and phone number. The leak of such information may cause identify theft, which can cause serious financial loss (Humphreys et al., 2010).

3. EVENT indicates the matters or activities that one involves anytime anywhere. For example, a tweet about taking a vacation implies that no one stays at home, which may result in burgalaries (Mao et al., 2011).

4. PREF refers to an individual's preference or hobbies. For example, a user posts a tweet about loving someone else and cheating on his/her partner may lead to blackmail.(Mansfield-Devine, 2015)

3.2.2 Definition of Private Rules

Therefore, given a word sequence, the judgment of privacy information can be summarized as a rule as follows:

$$\exists PERSON \wedge (\exists TRAIT \vee \exists PREF \vee \exists EVENT) \tag{3.1}$$

In accordance with the privacy rule above, when a post on OSNs involves the PERSON and TRAIT or PREF or EVENT entities, this post will be identified as a private post. A post having only TRAIT, PREF, and EVENT can not be considered as a private post because it not includes the user information, which is not suitable for the definition of privacy.

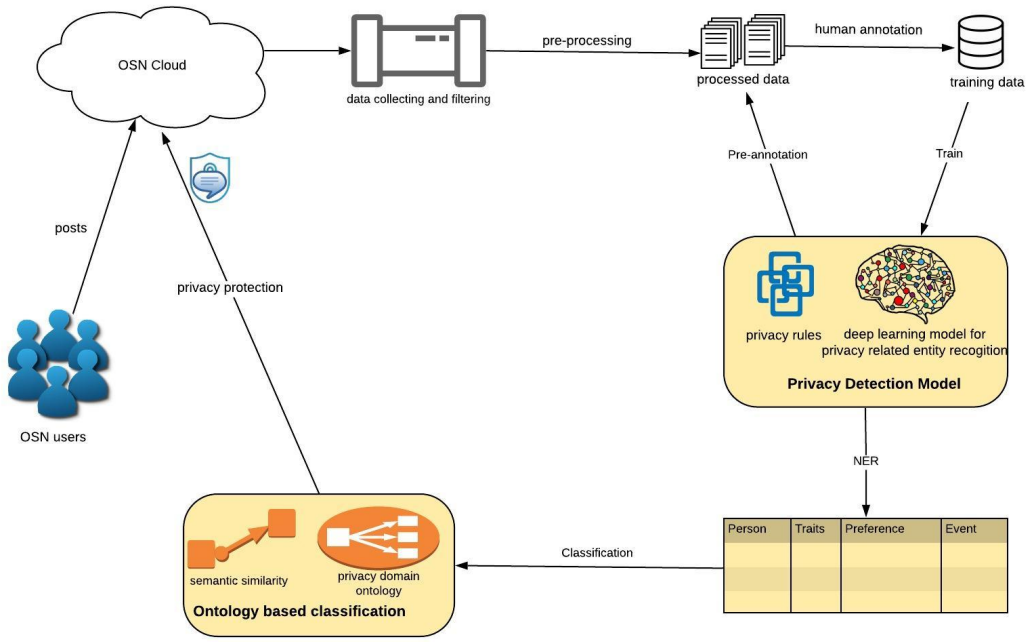

Figure 3.1: Automated Hybrid Privacy Detection Approach

3.3 Automated Hybrid Privacy Detection Approach

After the formal definition of individual privacy information on OSNs. The proposed automated hybrid privacy information detection approach is demonstrated in Fig 3.1. There are two key parts in the proposed framework, i.e., the privacy-related entities recognition and further classification based on ontology models. In the former part, a deep learning model is trained to detect the four types of entities that potentially cause privacy leakage. While, in the latter part, a privacy ontology model is developed based on the analysis of messages posted by the OSNs users.

The rationale of utilizing a hybrid model of both deep learning and ontology-based classification is clarified as follows: ontology-only approach classifies privacy information merely based on the domain-specific vocabulary of terms or concepts, which have to be systematically defined (Noy, McGuinness et al., 2001). Deep learning based NER does not necessarily require a lexicon or domain-specific terms, but it is

difficult to control the classification of specific terms as deep learning models turn out to be a "black-box". By considering the factors mentioned above, a combination of deep learning and ontology model has been adopted.

The details of the deep learning based detection approach introduced in the following subsections.

3.4 Deep Learning Based Detection approach

From Fig. 3.1 on the preceding page, users keep posting messages to the OSNs hosting in the cloud. Such raw, unstructured and public data can be obtained through crawlers or APIs provided by the OSNs. For example, Twitter allows developers to search public tweets if the proposed project is approved, and over 1 million tweets can be obtained every day through Twitter APIs. Given the context of privacy detection, the potential privacy-related data should be filtered and downloaded. The pre-processing step is conducted based on specific rules, such as removing messages which are advertisements or spams, removing meaningless words and characters and parsing word sequences to tokens. The processed data are supposed to be further enriched by running through the pre-annotation if a privacy detection NER model is available.

Next, human involvement, i.e., manual annotation, are required. Specifically, according to the aforementioned definition of privacy information, it is essential to recognise the privacy-related entities, i.e., PERSON, TRAIT, PREF, and EVENT. The annotation step also aims to figure out these four types of entities from the processed data. The annotated dataset is then fed into the deep learning model, i.e., the Bi-LSTM CRF model, to train it. If the privacy-related entities have been extracted from the deep learning-based model, then the aforementioned defined privacy rules will be applied to judge whether the tweet reveals private information.

The privacy detection model consists of two components, i.e., a deep learning

model for privacy-related entities recognition and privacy definition rules. For any messages posted by the OSNs users, the proposed model is capable of judging whether the message is privacy-related or not. Moreover, provided privacy rules are properly defined, the privacy detection model can also explain the reason why the message is potentially privacy-related. Using only a single deep learning model for private messages classification definitely reduces the capability of justification.

The privacy-related entities recognition plays an important role in the entire approach. There are two major aspects affecting the performance of a NER model, i.e., the annotation approach and the algorithm.

3.4.1 Outside-Inside-Beginning Annotation Approach

Annotation is a basic problem that we often encounter in order to fulfill an NER task. In the step of sequence annotation in NER, each element of a sequence is needed to be assigned a label.

Generally, there are two kinds of annotation approaches in NER (J.-D. Kim, Ohta, Tsuruoka, Tateisi & Collier, 2004):

1. Raw labelling: Each element in a named entity is supposed to be assigned a label.

2. Joint segmentation and labelling: It means that all segments in a named entity will be annotated with the same label. For example, given one sentence "Last month, Rohan Ford had a meeting", it includes a named entity regarding the type of person: Rohan Ford. Regarding the joint segmentation and labelling approach, the label "Person" is assigned to the entire phrase "Rohan Ford" instead of annotating the two words separately.

With regard to the annotation approaches, Outside-Inside-Beginning (BIO) encoding scheme is a standard method which can solve the joint segmentation problem in labelling

sequences. It is utilised to tag entities in a large number of NER tasks by transforming them into raw labelling problem (J.-D. Kim et al., 2004). Specifically, 'B-' is used as a prefix of an entity, implying the beginning of an entity; the prefix 'I-' tags other characters indicating the tag is inside of an entity, and 'O' is used for characters which do not belong to any pre-defined entities.

For example, privacy-related entities fall into BIO scheme are normally annotated as follows:

I B-PERSON

watch B-EVENT

a I-EVENT

movie I-EVENT

with O

Christine B-PERSON

3.4.2 NER Algorithm

In this thesis, the Bi-LSTM CRF model has been employed for privacy-related entities recognition in our model, as it is capable of achieving more promising results compared with those of other classic algorithms when being applied to NER (Lample et al., 2016). The Bi-LSTM model can learn long-range dependency due to the structure of the cell in the hidden layers. Moreover, it can adjust the impact of previous states on the current states through the forget gate, the input gate and the output gate in the cell (Graves & Schmidhuber, 2005).

However, it lacks the feature of analysis on the sentence level, which can be solved by the CRF layer. It is able to consider contextual conditions to make global optimal predictions. Combining the Bi-LSTM and CRF together can label sequence effectively when it ensures extraction of contextual features (Huang et al., 2015).

3.4.3 Word Representations

To create the input vectors of the Bi-LSTM CRF model, we adopt the word embedding approach. Word embedding is defined as multi-dimensional word vectors, generated by semantic vector space models which use vectors to represent each word. It is one of the most widely used techniques to represent the vocabulary in the Natural Language Processing tasks (Pennington, Socher & Manning, 2014). Moreover, word embedding is good at capturing the content of a word in a document, identifying semantic and syntactic similarity and relationships with other words, which can enrich our training data.

However, because of the fact that there is no publicly available and large word vector corpus of privacy information, multiple large scale external resources such as Wikipedia are necessary for building word embedding algorithms. Word2Vec and Global Vectors for Word Representation (GloVe) model are representative examples of word embedding approaches generated from external sources, they treat each word in the corpus as an atomic entity. However, most word embedding approaches like Word2Vec exhibit a disadvantage, which is the lack of co-occurrence between words. However, the GloVe approach trains global word-word co-occurrence counts which fills this gap, outperforming other current word representation approaches in common NER tasks (Pennington et al., 2014). Therefore, we utilise the Glove word embedding in privacy-related entities extraction.

Moreover, apart from the GloVe word embedding approach, character embeddings are also utilised in the NER model in this thesis. It aims to capture the morphological and orthographic privacy entity information because spelling evidence can help to extract an entity (Lakshmi et al., 2016). The model utilises a forward and backward LSTM to achieve a representation of each word from the concatenation of word embedding and corresponding character embeddings. After that, morphological information and the

contextual impact can be both achieved by the combination of character embedding and GloVe word embedding.

3.4.4 NER Based on Bi-LSTM model

Regarding the above, the Bi-LSTM model can maintain the sequential input information and assume labels independently. However, it does not consider the contextual conditions to assign the word label (Konkol & Konopík, 2013). For example, I-EVENT can not follow the B-PERSON. The CRF layer can assist the Bi-LSTM model with the contextual constraints function, which can bring benefits when considering the correlations between labels (Konkol & Konopík, 2013). Consequently, the Bi-LSTM CRF model is used widely in the NER task recently, which can obtain an outstanding result. By combining the strength of Bi-LSTM CRF model with word representation, we construct the privacy entities recognition model in Fig. 3.2 below.

From Fig. 3.2, l and r represent the left and right direction processing sequence of the Bi-LSTM encoder respectively. c represents the combination of the two direction vector yields of a word (Huang et al., 2015). The Bi-LSTM works as an encoder and CRF layer are used to decode the labels.

Given an input sequence $M = (m_1, m_2, , m_n)$, and the output matrix of Bi-LSTM model is P (Huang et al., 2015). $P_{i,j}$ can represent the score of label j assigned to word i. Consequently, the score of prediction sequence $O = (o_1, o_2, , o_n)$ can be calculated as follows:

$$s(M, o) = \sum_{i=0}^{n} A_{o_i, o_{i+1}} + \sum_{i=1}^{n} P_{i, o_i} \tag{3.2}$$

where A is the transition scores matrix, and $A_{i,j}$ can represent the transition score of label i to label j (Huang et al., 2015). Then for all the possible labelling sequences O_M, the conditional probability in the CRF layer is modified and the final label o^* can be

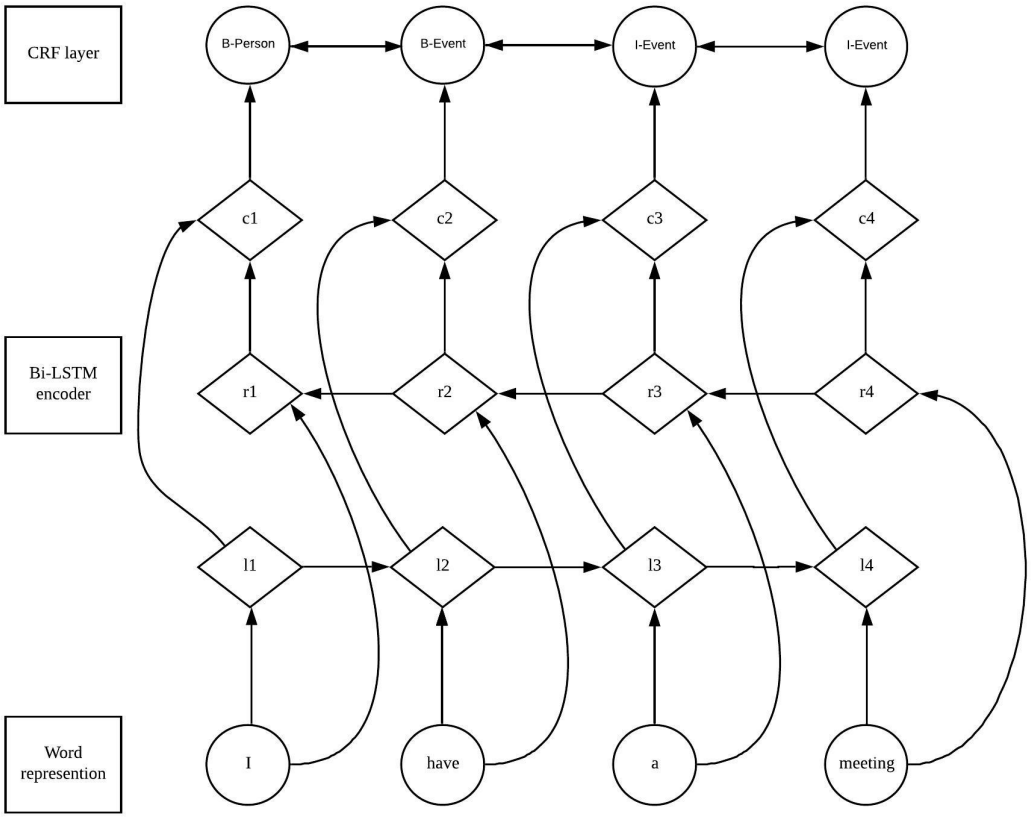

Figure 3.2: Bi-LSTM CRF Model in NER

calculated.

$$p(o|M) = \frac{e^{s(M,o)}}{\sum_{\tilde{o}\epsilon O_M} e^{s(M,\tilde{o})}} \tag{3.3}$$

$$o^* = \arg\max_{\tilde{o}\epsilon O_M} s(M, \widetilde{o}) \tag{3.4}$$

From Fig. 3.2 and equations above, the highest score of each word in a sentence can be marked by the CRF layer encoder and then its label can be determined. In comparison with the Bi-LSTM model, the strength of the combined model structure can bring the contextual conditions function, which considers the correlations between labels (Huang et al., 2015). The Bi-LSTM networks is utilised as the encoder while the

CRF layer is utilised to decode the suitable entity label of a word sequence.

3.5 Experiments

Two experiments have been conducted to evaluate the proposed privacy detection approach in this chapter. The first experiment aims to train a privacy-related entities recognition model using the Bi-LSTM CRF model with the word representation approach. It includes data collection (Subsection 3.5.1), data pre-processing (Subsection 3.5.2), evaluation metrics (Subsection 3.5.3), parameter tuning (Subsection 3.5.4) and GloVe word embedding (Subsection 3.5.5) to obtain the best performance. Then the second experiment evaluates some case studies to further demonstrate the effectiveness of the proposed privacy detection model.

3.5.1 Data Description

Twitter [1] is one of the largest Online Social Media, which enables users to conduct online social activities, i.e., the distribution of any ideas or information. In Twitter, the messages that are posted and interacted by users are known as “tweets”.

Twitter provides APIs, allowing developers to search and store Twitter provides APIs, allowing developers to search and store tweets. About over 1 million tweets can be obtained every day through the Twitter API (Cai et al., 2010). Therefore, we utilise Twitter API to collect 18000 tweets by searching for some terms which potentially result in privacy leakage, such as pronouns, sensitive words, plans, etc. For example, some tweets containing pronouns are chosen such as "I", "we", "she", "he", etc. Some verbs which can be indicators of topical relevant tweets, e.g., eating, shopping, as well as some nouns, e.g., concert, are frequently mentioned in event-related entities.

[1] https://twitter.com/

Additionally, some words are significant indicators of privacy-related entities, which can imply the user is leaking his/her personal sensitive information.

3.5.2 Data Pre-processing

After data collection is carried out through Twitter API, due to the purpose of the task of recognising privacy-related entities in tweets, the data pre-processing step is essential. The reason is that the highly unstructured data complexity of tweets makes it difficult to process data correctly, e.g., some noisy texts do not add value to the results (Medlock, 2006). So we take some effective pre-processing steps to eliminate unnecessary elements in this thesis.

There are several steps we can take to clean, normalise and tokenise tweets as follows:

1. Data cleaning: There are some non-textual data such as URL, # tags and special characters. The reason is that we found a large number of the tweets containing these characters are advertisements or spam.

2. Removing stop words: In the process of the training model, there are some terms for the results are invalid, such as some short common words, e.g., "a", "to", and the punctuation marks. These words that are not expected to be analysed in the texts are called stop words (Gomez-Hidalgo et al., 2010). Stop words contributes no value to the recognition results and are required to be removed.

3. Spelling Correction: In tweets, substitution and splitting steps are taken because of spelling correction. If the words in tweets are incorrect, then they will be substituted with the correct ones. Splitting means splitting those merged words, e.g., "I'm" will be split into "I am". The two steps make it easier to recognise privacy-related entities, especially person-related entities.

4. Tokenisation: In English, spaces are natural boundaries between words, and punctuation marks are separators between sentences. Therefore, the text could be split through Regular Expression using spaces and punctuation.

3.5.3 Evaluation Metrics

In this experiment, after data preprocessing, we leverage three traditional evaluation metrics serving NER tasks: Precision, Recall, and F1-score (Gomez-Hidalgo et al., 2010). When the three values are higher, the experimental results are more outstanding.

There are four possible outcomes in the results as below:

1. An entity will be recognised as a privacy-related entity when it is truly a privacy-related entity (True Positive, TP);

2. An entity will be recognised as a privacy-related entity when it is actually a non-privacy-related entity (False Positive, FP);

3. An entity will be recognised as a non-privacy-related entity when it is actually a privacy-related entity (False Negative, FN);

4. An entity will be recognised as a non-privacy-related entity and it is truly a non-privacy-related entity (True Negative, TN).

Based on the above four possible results, Precision, Recall, and F1-score can be defined as follows:

1. Precision: the percentage that privacy-related entities can be labelled correctly among all the entities which are labelled privately in the test dataset.

$$Precision = TP/(TP + FP) \tag{3.5}$$

2. Recall: the percentage that privacy-related entities can be labelled correctly among all the actual privacy entities in the test dataset.

$$Recall = TP/(TP + FN) \tag{3.6}$$

3. F1-score: the weighted average of precision and recall, which takes both the two measures into account. It is also considered as the comprehensive evaluation value of the results.

$$F1 = (2 \times Precision \times Recall)/(Precision + Recall) \tag{3.7}$$

3.5.4 Parameters Tuning

In the Bi-LSTM CRF model, there are a variety of parameters can be tuned (Qin & Zeng, 2018). Among them, the dropout value is similarly set as 0.5. Dropout can be used as an approach to solving the problem of over-fitting when training deep neural networks (Krizhevsky, Sutskever & Hinton, 2012). When a complex feedforward neural network is trained based on a small data set, the problem of over-fitting can easily occur. The performance of the training neural network is difficult to be improved when the model is over-fitting.

However, over-fitting can be significantly reduced by ignoring half of the feature detectors in each training batch of the neural network, i.e., assigning half of the hidden layer nodes with zero value. This approach will reduce the interaction between hidden layer nodes (Hinton, Srivastava, Krizhevsky, Sutskever & Salakhutdinov, 2012). Consequently, the performance of the neural network can be improved by setting the dropout value as 0.5.

In the parameter tuning step, we keep trying different combinations of those parameters and record the corresponding results according to the value of F1-score based on the training data set.

Eventually, we can obtain a result of the optimised parameter setting, as demonstrated in Tab. 3.1.

Table 3.1: The Optimised Parameter Settings of The Bi-LSTM CRF Model

Parameter	Setting	Description
word embedding	TRUE	Using word embedding
use_char	TRUE	Using character embedding
char_lstm_size	25	Word tagger LSTM output dimensions
use_crf	TRUE	Using CRF
dropout	0.5	input dropout rate
num_labels	10	Number of entity labels
fc_dim	100	Output fully-connected layer size
word_LSTM_size	100	Character LSTM feature extractor output dimensions
char_embedding_dim	25	Character embedding dimensions

3.5.5 Word Embedding Trained from Different External Sources

When applying the word embedding approach, different external sources where pre-trained word vectors come from can impose influence on the quality of experimental results (Pennington et al., 2014). To validate whether the domain external sources will influence the experimental results, we use pre-trained word vectors come from the non-domain and domain sources in Glove:

1. Pre-trained word vectors which contain 6B tokens and 400K vocabulary come from Wikipedia 2014 and English Gigaword Fifth Edition.

2. Pre-trained word vectors that contain 2B tokens and 1.2M vocabulary come from Twitter.

We take the experiments of privacy-related NER based on the two external sources above and record the corresponding results. As demonstrated in Tab. 3.2, the performance of them can be shown:

Table 3.2: Results of Different External Sources of Word Embedding

Pre_data sets	Precision	Recall	F1-score
Wikipedia	0.88	0.91	0.90
Twitter	0.90	0.92	0.91

From the result in Tab. 3.2, we can see if the word vectors of the training dataset are pre-trained on domain source corpora, i.e., Twitter, it can get better performance than the non-domain source corpora, i.e., Wikipedia and English Gigaword, regarding the value of Precision and Recall, but the difference is tiny. We can draw the conclusion that word vectors of the training data set pre-trained on the domain source corpora can help to make more precise privacy-related judgments instead of only detecting more suspect privacy-related entities.

In addition, the value of word embedding dimensions may also influence the result performance. Similarly, we set the word embedding dimensions from 25 to 200. Tab. 3.3 demonstrates the experimental results of the different word embedding dimensions.

Table 3.3: Results of Different Word Embedding Dimensions

Size	Precision	Recall	F1-score
25	0.88	0.81	0.84
50	0.91	0.90	0.90
100	0.90	0.92	0.91
200	0.93	0.94	0.93

From the results demonstrated in Tab. 3.3, the highest F1-score is 0.93 when the value of word embedding size is 200. Consequently, we can use 200 as the value of word embedding dimensions, which is also commonly used in a large number of other NER tasks as the best word embedding dimensions size setting, e.g., the size of word embedding dimensions is set to 200 (Yang, Macdonald & Ounis, 2018).

3.5.6 Experimental Results

After the steps of data collection, data pre-processing, parameter tuning and GloVe word embedding, two experiments can be conducted to evaluate the proposed deep learning based NER model by using the real-world dataset collected from Twitter API.

- Experiment 1

In Experiment 1, a privacy detection model based on Bi-LSTM CRF with word representation is trained to recognise the privacy-related entities after parameters tuning. Through it, the users can be alerted before potential privacy leakage occurs.

According to the definition of privacy and BIO encoding scheme mentioned previously, nine tags have been defined, i.e., 'B-PERSON', 'I-PERSON', 'B-TRAIT', 'I-TRAIT', 'B-PERF', 'I-PERF', 'B-EVENT', 'I-EVENT' and 'O'. Around 18000 tweets have been annotated manually by applying these nine tags.

After 50 epochs training in the deep-learning based model, with the parameters tuning and word embedding setting which can perform the best results, the detailed four types entity recognition performance of the deep learning-based model is demonstrated in Tab. 3.4.

Table 3.4: Performance of Privacy-Related Entities Recognition

Entity	Precision	Recall	F1-score
PREF	0.89	0.89	0.89
TRAIT	0.87	0.99	0.93
PERSON	0.90	0.87	0.88
EVENT	0.97	0.97	0.97
Avg/Total	0.93	0.94	0.93

From the results of the four types of entities, we can find the four types of privacy-related entities recognition both achieve high F1-scores. However, the entity type of "PREFERENCE" is the most difficult to be recognised because identifying the border of chunks of this type can be a little ambiguous. Some preference descriptions in tweets

can be as long as eight words, for example, *an expensive book came from my favourite friend*. That is because the contextual clues cannot be informative enough to recognise this type.

- Experiment 2: Case Study

In this experiment, we further demonstrate the effectiveness of the proposed privacy detection model by selecting three tweets posted recently and analysing the results produced by the deep learning-based model. The results contain the labels produced by the model, and corresponding explanations are based on the privacy rules we defined previously.

Case 1: *Adam and I are having lunch tomorrow*.

Results: Adam (B-PERSON) and I (B-PERSON) are having (B-EVENT) lunch (I-EVENT) tomorrow.

Explanation: Based on the privacy rules, this tweet is privacy-related since it mentions both PERSON and EVENT entities.

Case 2: *Watching a movie is a good way to relax!*

Results: Watching (B-EVENT) a (I-EVENT) movie (I-EVENT) is a good way to relax!

Explanation: Based on the privacy rules, this tweet is just a simple statement regarding *Watching a movie*, which is not a private one because it does not contain the PERSON entity.

Case 3: *My son is crazy about coke*.

Results: My (B-PERSON) son (I-PERSON) is crazy about coke (B-PREF).

Explanation: Based on the privacy rules, this tweet talks about PERSON and PREF, it is privacy-related.

3.6 Summary

In this chapter, we propose a hybrid privacy detection approach for individual OSNs users after giving formal definitions. The objective is to protect users from potential privacy leakage before posting any messages. The proposed framework explains the process of data collection, data processing, model training and how it works. Both privacy rules and the Bi-LSTM CRF model are leveraged in the privacy detection model. Thus, the proposed model is equipped with the capability of explaining both the detection and the results.

Then through a real-world training dataset extracted from Twitter API, data preprocessing and parameter tuning are conducted to improve the model performance. The first experiment uses the evaluation metrics (Precision, Recall, and F1-score) to demonstrate the high accuracy of the model we formed. The second experiment gives several case studies to detect and explain results according to the privacy rules we defined previously. Through the two experiments, we can deduce that the proposed deep learning based NER model has significant performance advantages.

Chapter 4

Ontology-based Privacy Information Classification

4.1 Introduction

Massive amounts of information are being published published to online social networks every day. Individual privacy-related information is also possibly disclosed unconsciously by the end users. Identifying privacy-related data and protecting the online social network users from privacy leakage is important for protecting these end users. In Chapter 3, in the context of OSNs' "user privacy", refers to a sequence of words, stating or implying any individual's personal information, preferences, events that he or she is involved in; privacy leakage describes a situation when an individual shares stories including private information with their contacts or even those they are not familiar with. Thus, it is necessary to develop a tool to detect and identify all the possible privacy-related information contained in any messages posted (Wang et al., 2011; Hasan et al., 2013). More importantly, the justifications of privacy-related information classification would be helpful for OSN users to avoid posting similar messages again.

In Chapter 3, we leverage a hybrid privacy classification approach, incorporating both deep learning and ontology models, for individual users of OSNs. The first part of the approach is a generic approach to privacy leakage detection for OSNs users. Based on our two experiments, its capability is considerably enhanced which means it is capable of capturing privacy-related entities after giving sufficient training.

However, two significant limitations remain. Firstly, it can only remind users regarding the possibility of privacy leakage. As a result, detailed leaking information in terms of what kind of leakage can occur is missing. However, people may pay attention to the categories of disclosure where there exists clear risk and may cause serious consequences to the user. Secondly, the privacy model is not conceptually modelled or presented. Ontology models conceptually reflect the domain-specific knowledge in the form of terms and demonstrate two apparent advantages, i.e., shareability and reusability (Zhao et al., 2009). Therefore, ontology-based privacy models can be easily extended and applied to different OSNs (Mitra et al., 2005). Moreover, because ontology organises the concepts in the form of taxonomy or hierarchy based on a pre-defined natural relationship, it is suitable to introduce ontology into the research of privacy-related information extraction and classification.

As the second part of our hybrid privacy detection approach, it can address the privacy leakage problem for individuals by effectively identifying privacy information and classifying into a detailed category. More specifically, deep learning based models are utilised to conduct the Name Entity Recognition (NER) and detect the pre-defined privacy-related entities. An ontology model is developed based on the analysis of the data collected from real-world users. The privacy ontology model will further classify the recognised entities into sub-classes according to the results carried out by the deep learning models discussed in this Chapter.

The remainder of this chapter is structured as follows. Firstly, the domain and scope of the privacy domain ontology are described (Section 4.2) as it is the first step to

build an ontology (Noy et al., 2001). Secondly, through scanning the NER results from the real-world dataset extracted from Twitter API, the representative keywords related to privacy are demonstrated (Section 4.3). Thirdly, the privacy domain ontology for OSNs can be presented in Section 4.4 after developing a class hierarchy. In Section 4.5, for classifying privacy-related entities into fine-grained categories, a semantic phrase similarity degree approach based on GloVe is presented. And finally, experiments are conducted to identify the effectiveness of the ontology-based classification approach (Subsection 4.6.2) and the results of privacy leaking are categorised in Subsection 4.6.3 to be discussed.

4.2 The Domain and Scope of The Privacy Domain Ontology

The defining domain and scope of "privacy" is the first step to build a private-related information ontology model (Noy et al., 2001). Naturally, the privacy information concepts describing different subclasses (class corresponds to the entity type in this thesis) of the four privacy-related entities will be fed into our ontology, given the privacy domain ontology we are going to build in this thesis. Specifically, the privacy domain ontology on OSNs includes:

1. The concepts of privacy ranges from general classes to hierarchical subcategories of specific subclasses.

2. A set of relationships between privacy classes that link concepts in a more complex way, implicit in the underlying hierarchy.

4.3 Privacy-related Keywords Extraction

Initially, it is significant to obtain a comprehensive list of privacy-related terms and concepts in order to form the hierarchy of privacy ontology. To construct an ontology, we extracted all the values of privacy-related entities recognised by the deep learning based model. Next, based on the word frequencies, representative keywords are selected as the major indicators of the subcategories of the entities. For example: we want to find some keywords regarding private events under the "Event" main class. Some verbs representative of private events, e.g. eating, shopping, etc., as well as some nouns, e.g. concert, meeting, journey, etc., are frequently mentioned in event-related entities. Additionally, some words are significant indicators of privacy-related entities, which can imply the user is leaking his/her "TRAIT" sensitive information.

Because the creation process of an ontology is an interactive process (Madsen et al., 2006), we searched the selected keywords in the dataset to find out the occurrence of these words and how important of them according to the term frequency (Noy et al., 2001). Through the interactive procedure, the terms and concepts of the privacy information ontology are finally determined. For example, because the keyword of "interview" frequently appears in the extracted entities, it turns out to be a keyword, representing the subclass of "Corporate Event".

Table 4.1 shows the representative keywords extracted from the collected data, which are utilised for building the privacy domain ontology on OSNs.

4.4 Privacy Ontology

Among the possible approaches in forming a class hierarchy, a top-down approach has been selected by considering the relationships between the privacy concepts in this thesis (Noy et al., 2001). The ontology hierarchy presents a tree structure, having most

Table 4.1: Corresponding Keywords with Classes and Subclasses

Class	*Subclass*	*Keywords*
Person	Individual	I
	Third Party	you, we, they, he, she, classmate, uncle
Preference	Item	book, chocolate, keyboard, tea
	Hobby	cosplay, paint, fishing, dancing, reading
	Specific Person	girlfriend, teacher
Event	Private Event	eat, shopping, concert, movie, exercise, spa
	Corporate Event	wedding, interview, meeting, conference, festival, party, parade, salon
	Journey	fly, holiday, travel, island, hotel, airport
Trait	Individual Identity	years-old, Auckland
	Linked Information	lawyer, female, gay, Christian, married, white, disable

general classes on top and specific associative classes connected with the general ones. For example, given "PERSON" as a superclass, "Individual" and "Third Party" can be the subclasses based on the recognised entity-value pairs. Three other superclasses, i.e., "TRAIT", "EVENT", and "PREFERENCE", are also included.

"TRAIT", i.e., sensitive personal information, which is defined and classified into two types from the usage of the Personally Identifiable Information (PII) in United Stated legal terms, i.e., distinguish identity and relating information (McCallister, Grance & Scarfone, 2010). Similarly, "TRAITS" can be classified into two subclasses as below:

1. Any information can be utilized to identify an individuals identity, e.g., birth date and hometown.

2. Any information can be linked to an individual, e.g., medical records, educational background and marital status.

Through this kind of classification approach, two subclasses of "Trait" are identified: "Individual identity" and "Linked information". For example, the date of birth can be recognised as "Individuals identity". Whereas, race, gender, sexual orientation,

marital status, religion, belief, and education background are categorised as "Linked information". Similarly, EVENT can be classified as a private event, corporate event, and journey in terms of the event is social or non-social. Among the subclasses of "Event", tweets about the journey is individually classified because we think users who reveal their journey plans will make them very vulnerable to robbery crimes. Corporate event means an event included lots of people, e.g., wedding ,meeting. Preference can be classified as a specific person, item and hobby according to the characteristic of the leaking hobby information. Therefore, the privacy domain ontology for OSNs is presented in Fig. 4.1:

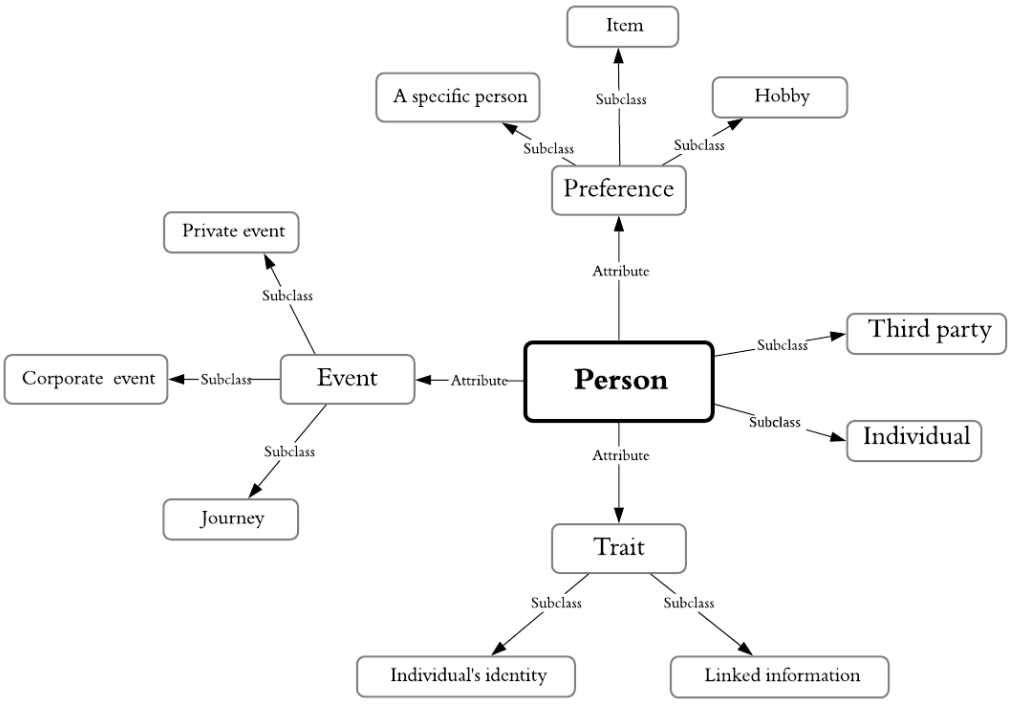

Figure 4.1: Privacy Ontology

4.5 Semantic Phrase Similarity Degree Based On GloVe

To classify the privacy-related entities into fine-grained subclasses, it is essential to take an approach in finding the subclass where the entities have the highest degree of belonging. Moreover, the highest belonging degree is decided by the highest similarity degree between the extracted entities and the representative terms. Comparing the similarity of word vectors or word embedding of each word can decide their similarity degree (Jiang & Conrath, 1997). Word embedding is defined as multi-dimensional representations of a word, which can be generated by semantic vector space models that use vectors to represent each word.

Consequently, in this section, we will propose a semantic phrase similarity degree approach based on semantic vector space models. It is divided into two steps. Firstly, a word semantics vector space model is chosen (Subsection 4.5.1). Secondly, the construction of the classification model will be described in Subsection 4.5.2 by taking "Event" entity as an example.

4.5.1 Word Semantics Vector Space Model

The first step is to choose a word semantics vector space model. A pre-trained statistical model (called "en_core_web_lg" in spacy) is used in this thesis, which is trained on blogs, news, and comments with GloVe trained vectors. Global Vectors for Word Representation (GloVe) is a state-of-art tool using word embedding techniques. Most word embedding approaches like Word2Vec exhibit a disadvantage, which is the lack of co-occurrence between words. Luckily, the GloVe approach trains global word-word co-occurrence counts which fills this gap, outperforming other current word embedding approaches in common word similarity tasks (Pennington et al., 2014) (Mihalcea, Corley, Strapparava et al., 2006). That is the reason we choose this word semantics vector space model to use in order to compare the similarity degree between words.

4.5.2 The Construction of The Classification Model

To explain the construction of the classification model, we make use of "Event" entities as an example. In Fig. 4.1, the event class consists of three subclasses, "Private Event", "Corporate Event" and "Journey". Moreover, in Table 4.1, there are 20 representative terms representing the subclasses associated with them. Each representative term can be represented as $term_i$, where i belongs to {0,1,2,...,19}. Among them, the value i of "Private Event":{0,1,2,...,5}, the value i of "Corporate Event":{6,7...,13}, and the value i of "Journey":{14,15...,19}. Similarly, each word in the extracted entities can be represented as $entity_j$.

Consequently, S_{ij} can represent the semantic similarity degree between each word in the extracted entities $entity_j$ and each representative term $term_i$:

$$S_{ij} = similarity(entity_j, term_i) \tag{4.1}$$

So the similarity of event-related entities and the event representative terms in the ontology can be calculated as follows:

$$S_{i(sum)} = \sum_{j=1}^{n} S_{ij} \tag{4.2}$$

After the calculation of all the similarity degrees, the subclass of the extracted event entities can be decided according to the maximum degree of $S_{i(sum)}$, which means the subclass which obtains the maximum $S_{i(sum)}$ is the corresponding subclass of the event-related entity:

$$\lambda = \max\left(S_0(sum), S_1(sum), ..., S_{19}(sum)\right) \tag{4.3}$$

Then the i which obtains the maximum degree of $S_{i(sum)}$ can be decided and the corresponding subclass can be deduced as below:

$$subclass = \begin{cases} PrivateEvent & i \in 0, 1, ..., 5 \\ CorporateEvent & i \in 6, 7, ..., 13 \\ Journey & i \in 14, 15, ..., 19 \end{cases}$$

Hence, by using the semantic information which the word embedding technique captures (Yaguinuma, Ferraz, Santos, Camargo & Nogueira, 2010), we can classify privacy-related information into the corresponding subclass in the privacy ontology like the “Event" procedure we list.

4.6 Experiments

Experiments have been conducted to evaluate the proposed hybrid privacy detection approach in this thesis. The experiment uses a real-world dataset that aims to classify fine-grained privacy-related entities. Moreover, we also focus on one interesting problem concerning private tweets: What type of personal private information is leaked most on OSNs?

4.6.1 Data Description

Twitter [1] is one of the largest Online Social Media platforms as a micro-blogging service, where a large amount of information is broadcast publicly by individual users. In Twitter, the information posted by end users are named as tweets. Twitter provides APIs, allowing developers to search and store tweets based on certain criteria.

Therefore, we use Twitter API to search for some terms which contain sensitive keywords, e.g., sensitive activities and plans, which may result in privacy information leakage and lead to negative consequences. Then we collect around 18000 tweets as a

[1] https://twitter.com/

testing dataset in this experiment. As we demonstrated before, the tweets will be conducted with the deep learning based NER approach and the ontology-based classification approach in this chapter, then the corresponding subclasses will be decided.

4.6.2 Evaluation

To evaluate the performance of the hybrid privacy detection approach, we manually annotate the subclasses of privacy leaking information. Moreover, the “ground truth” is prepared by allowing the users themselves to provide opinions on whether they leak the privacy and what types of private information they are leaking on tweets. For example, a tweet “I watch a movie.", has “I" annotated as “Individual" and “watch a movie" annotated as “Private Event".

Four traditional measures are utilised to evaluate the performance of the proposed hybrid approach:

1. Accuracy: the fraction of correct classification in the testing data set;
2. Precision: the fraction of correct classification among all results are classified in this subclass in the testing data set;
3. Recall: the fraction of correct classification among all actual results belong to this subclass in the testing data set;
4. F1-value: the harmonic value of precision and recall, which is a balance measurement.

We emphasize the accuracy metric among the traditional measures. In our detection approach, it is more important to be accurate to detect privacy information.

Table 4.2: Performance of Hybrid Privacy Information Classification

Class	*Subclass*	*Accuracy*
Person	Individual	0.94
	Third Party	0.85
Preference	Item	0.82
	Hobby	0.81
	Specific Person	0.76
Event	Private Event	0.74
	Corporate Event	0.76
	Journey	0.77
Trait	Individual Identity	0.62
	Linked Information	0.64

4.6.3 Experimental Results

Our hybrid privacy classification approach utilises a deep learning based NER approach and an ontology-based approach to perform classification of specific privacy information. After the NER, we build a privacy domain ontology and use the ontology vocabulary for performing the semantic phrase similarity degree calculation with the extracted privacy-related entities.

We evaluate our hybrid approach on the testing dataset and get an excellent accuracy performance, as shown in Table 4.2, with high accuracy in each type. We believe this accuracy is high enough to indicate the effectiveness of our automated detection and classification approach. However, we observe the accuracy of categories under the “TRAIT" entity is much lower than other types of entities. The extraction result of the “TRAIT" entity is lower than other types of entity in NER and a trait-related privacy entity will be classified to categories under other types of entities. That is why the accuracy value of classification of “TRAIT" entity is lower than other entities.

4.6.4 Privacy Leaking Information Categorization

After the whole automated hybrid privacy information detection approach, all the types of private information leaks on the 18000 testing dataset can be extracted. Then we plot the distribution of the results of the privacy information types of the testing dataset in Fig. 4.2, which includes all the subclasses under the “TRAIT", “PREF", and “EVENT" entities. According to the privacy rule in this detection approach, each tweet with privacy-related entities which is identified as the private tweet contains the “Person" entity, so the distribution of the “Person" entity is not necessary to be analysed in the aspect of its categorization.

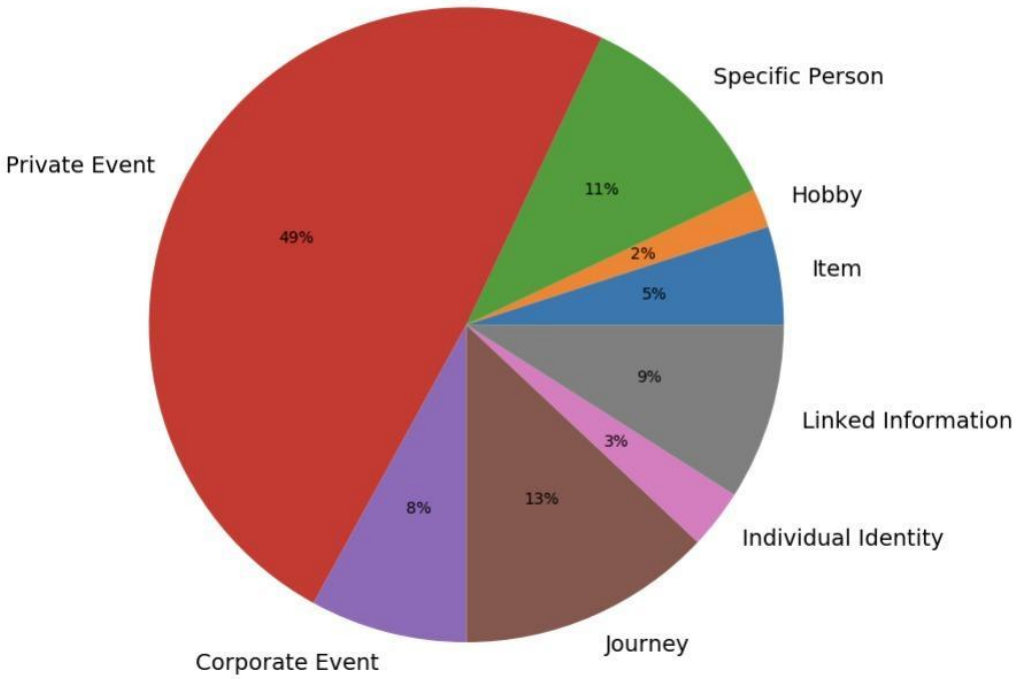

Figure 4.2: Distribution of Different Types of Privacy Information Leaking

In this section, we also explore the question of what type of sensitive information users leak the most. We counted the percentage of the eight types of privacy information in the 18000 testing dataset. The results are shown in Fig. 4.2. From Fig. 4.2, we can see the most leaking information is the “Event" entity, where “Private Event" counts the most in it. Additionally, “Journey" under the “Event" and “Specific person" under the “Preference" are also leaked a lot in tweets, which means the information about all subclasses under the entity of “Event" both is leaked a lot on OSNs. From the results, we suggest that OSNs users should exercise a little more restraint about posting relevant tweets about these types. On the other hand, people are more conservative about “Individual Identity" and “Hobby" information because the privacy leaks (as the percentage) of the two types is much smaller than other types.

4.7 Summary

In this chapter, we have proposed and evaluated an ontology-based classification approach to further classify privacy-related information to specific categories as the second part of our proposed hybrid privacy detection approach. Through characterising the nature of privacy information leaking on OSNs, a privacy domain ontology is built to automatically classify fine-grained privacy information, i.e., nine sub-types of private leaking. The ontology-based approach calculated the semantic similarity degree between entities extracted from the deep learning model and the representative terms in the ontology. We evaluated the result with the accuracy value, which demonstrates it gains a considerable performance. Moreover, what specific types of personal private information users are leaking on OSNs can be understood.

Chapter 5

Conclusions and Future Work

Privacy information detection is a hot topic in Online Social Media research. With the proliferation and popularisation of the World Wide Web, Online Social Networks have become of one of the significant channels for social interactions and communications. OSNs provide great convenience for the users, but these online social platforms also carry potential risks, such as privacy leakage. A vast amount of private information can be accessed publicly through OSNs, such as preferences, email address, marital status, hometown, activities attended, etc., which may lead to severe security issues, e.g., phishing, identity theft, romance fraud, and cyberstalking.

Therefore, it is significant to explore a useful and practical approach to protect OSN users from privacy information disclosure. Users should be alerted before posting any privacy-related messages to the public. More importantly, the justifications of privacy-related information classification would be helpful for OSN users to avoid posting similar messages again. However, researches about the privacy detection approach for protecting individual OSNs users are rare. There are two reasons. Firstly, most of the research only considers the perspective of corporations. Secondly, there are no clear definitions about individual privacy information on OSNs, and the related research only pays attention to several specific categories of privacy information.

In this thesis, we have proposed a hybrid privacy detection approach to protect individual users of OSNs; this incorporates both deep learning and ontology models. In this chapter, the major contributions of the thesis are summarised, and the future research paths are outlined.

5.1 Summary of Major Contributions

There are three major contributions we make in this thesis.

Firstly, we give a clear definition of individual privacy information on online social media. It is defined as word sequences implying the personal information, likes and dislikes and the events the user is involved in. Moreover, we define the private rules to judge whether the tweet is private or not according to the results of extracting these word sequences.

Secondly, we present a hybrid privacy information detection approach for individual OSNs users. The objective is to protect OSNs' users from potential privacy leakage before posting any messages about themselves. It comprises a deep learning based NER approach and an ontology-based classification approach. The proposed deep learning approach explains the process of data collection, data pre-processing, human annotation, model training and how it works to protect users. Both privacy rules and Bi-LSTM CRF model with word embedding are leveraged in the privacy detection model. Thus, the proposed model is equipped with the capability of both detection and explaining the results with considerable success.

Thirdly, the second part of our proposed hybrid privacy detection approach is ontology-based classification. Because of the research gap in privacy information ontology on OSNs and the characteristics of ontology models, it is essential to build a privacy domain ontology. Then we propose a semantic similarity degree approach to classify privacy related entities extracted from NER results into fine-grained privacy

information, which also performed very well.

5.2 Future Work

This research is still preliminary and it can be extended by investigating the following research paths.

Firstly, we intend to utilise a larger training data set for performance evaluation and improve the performance of the privacy-related entities recognition.

Secondly, instead of only recognising privacy-related entities from multiple users, in the future, all privacy information revealing by a user can be collected by the model and used to protect the user.

Thirdly, different tweets are associated with different degrees of privacy leakage. In the thesis, we demonstrate what types of privacy-related entities OSNs users reveal. However, we do not analyse the detailed privacy leakage degree of users. In the future, we plan to evaluate the privacy leakage degree and explore the insights based on predictive results.

References

Acquisti, A. & Gross, R. (2006). Imagined communities: Awareness, information sharing, and privacy on the facebook. In *International workshop on privacy enhancing technologies* (pp. 36–58).

Aldhafferi, N., Watson, C. & Sajeev, A. (2013). Personal information privacy settings of online social networks and their suitability for mobile internet devices. *arXiv preprint arXiv:1305.2770*.

Aronson, A. R. (2001). Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the amia symposium* (p. 17).

Batra, A., Sidhu, K. & Sharma, S. (2018). Characteristics of women whatsapp users and use pattern. *Journal of Education, Society and Behavioural Science*, 1–7.

Bhagat, S., Cormode, G., Srivastava, D. & Krishnamurthy, B. (2010). Prediction promotes privacy in dynamic social networks. In *Wosn.*

Cai, L., Lu, C. & Zhang, C. (2010). Privacy domain-specific ontology building and consistency analysis. In *2010 international conference on internet technology and applications* (pp. 1–6).

Collobert, R. & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on machine learning* (pp. 160–167).

Cormode, G. & Krishnamurthy, B. (2008). Key differences between web 1.0 and web 2.0. *First Monday*, *13*(6).

Coyle, C. L. & Vaughn, H. (2008). Social networking: Communication revolution or evolution? *Bell Labs Technical Journal*, *13*(2), 13–17.

Derzakarian, A. (2017). The dark side of social media romance: Civil recourse for catfish victims. *Loy. LAL Rev.*, *50*, 741.

Dewan, P., Kashyap, A. & Kumaraguru, P. (2014). Analyzing social and stylometric features to identify spear phishing emails. In *2014 apwg symposium on electronic crime research (ecrime)* (pp. 1–13).

Dwyer, C., Hiltz, S. & Passerini, K. (2007). Trust and privacy concern within social networking sites: A comparison of facebook and myspace. *AMCIS 2007 proceedings*, 339.

Farmakiotou, D., Karkaletsis, V., Koutsias, J., Sigletos, G., Spyropoulos, C. D. & Stamatopoulos, P. (2000). Rule-based named entity recognition for greek financial texts. In *Proceedings of the workshop on computational lexicography and multimedia dictionaries (comlex 2000)* (pp. 75–78).

Francia III, G. A., Hutchinson, F. S. & Francia, X. P. (2015). Privacy, security, and identity theft protection: Advances and trends. In *Handbook of research on emerging developments in data privacy* (pp. 133–143). IGI Global.

Galizia, S. (2006). Wsto: a classification-based ontology for managing trust in semantic web services. In *European semantic web conference* (pp. 697–711).

Ge, J., Peng, J. & Chen, Z. (2014). Your privacy information are leaking when you surfing on the social networks: A survey of the degree of online self-disclosure (dosd). In *2014 ieee 13th international conference on cognitive informatics and cognitive computing* (pp. 329–336).

Gehrke, J., Lui, E. & Pass, R. (2011). Towards privacy for social networks: A zero-knowledge based definition of privacy. In *Theory of cryptography conference* (pp. 432–449).

Gharib, M., Giorgini, P. & Mylopoulos, J. (2017). Towards an ontology for privacy requirements via a systematic literature review. In *International conference on conceptual modeling* (pp. 193–208).

Gomez-Hidalgo, J. M., Martin-Abreu, J. M., Nieves, J., Santos, I., Brezo, F. & Bringas, P. G. (2010). Data leak prevention through named entity recognition. In *Social computing (socialcom), 2010 ieee second international conference on* (pp. 1129–1134).

Graves, A. & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, *18*(5-6), 602–610.

Hasan, O., Habegger, B., Brunie, L., Bennani, N. & Damiani, E. (2013). A discussion of privacy challenges in user profiling with big data techniques: The eexcess use case. In *Big data (bigdata congress), 2013 ieee international congress on* (pp. 25–30).

Henderson, S. C. & Snyder, C. A. (1999). Personal information privacy: implications for mis managers. *Information & Management*, *36*(4), 213–220.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Huang, Z., Xu, W. & Yu, K. (2015). Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Humphreys, L., Gill, P. & Krishnamurthy, B. (2010). How much is too much? privacy issues on twitter. In *Conference of international communication association, singapore.*

Jiang, J. J. & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.

Kemp, S. (2018). Digital in 2018: World's internet users pass the 4 billion mark. *We are social*.

Kim, A., Hoffman, L. J. & Martin, C. D. (2002). Building privacy into the semantic web: An ontology needed now. In *Proc. of semantic web workshop, hawaii, usa.*

Kim, J.-D., Ohta, T., Tsuruoka, Y., Tateisi, Y. & Collier, N. (2004). Introduction to the bio-entity recognition task at jnlpba. In *Proceedings of the international joint*

workshop on natural language processing in biomedicine and its applications (pp. 70–75).

Konkol, M. & Konopík, M. (2013). Crf-based czech named entity recognizer and consolidation of czech ner research. In *International conference on text, speech and dialogue* (pp. 153–160).

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Lafferty, J., McCallum, A. & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Lakshmi, G., Panicker, J. R. & Meera, M. (2016). Named entity recognition in malayalam using fuzzy support vector machine. In *2016 international conference on information science (icis)* (pp. 201–206).

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. & Dyer, C. (2016). Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436.

LeFebvre, L., Blackburn, K. & Brody, N. (2015). Navigating romantic relationships on facebook: Extending the relationship dissolution model to social networking environments. *Journal of Social and Personal Relationships*, *32*(1), 78–98.

Li, J. (2014). Data protection in healthcare social networks. *IEEE software*, *31*(1), 46–53.

Madsen, M. et al. (2006). A health information privacy ontology: toward decision support for compliance assessment. *HIC 2006 and HINZ 2006: Proceedings*, 184.

Mansfield-Devine, S. (2015). The ashley madison affair. *Network Security*, *2015*(9), 8–16.

Mao, H., Shuai, X. & Kapadia, A. (2011). Loose tweets: an analysis of privacy leaks on twitter. In *Proceedings of the 10th annual acm workshop on privacy in the electronic society* (pp. 1–12).

McCallister, E., Grance, T. & Scarfone, K. A. (2010). Sp 800-122. guide to protecting the confidentiality of personally identifiable information (pii).

McFarlane, L. & Bocij, P. (2003). An exploration of predatory behaviour in cyberspace: Towards a typology of cyberstalkers. *First monday*, *8*(9).

Medlock, B. (2006). An introduction to nlp-based textual anonymisation. In *Lrec* (pp. 1051–1056).

Mihalcea, R., Corley, C., Strapparava, C. et al. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *Aaai* (Vol. 6, pp. 775–780).

Mitra, P., Liu, P. & Pan, C.-C. (2005). Privacy-preserving ontology matching. In *Aaai workshop on context and ontologies.*

Nadeau, D. & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, *30*(1), 3–26.

Naslund, J. A., Grande, S. W., Aschbrenner, K. A. & Elwyn, G. (2014). Naturally occurring peer support through social media: the experiences of individuals with severe mental illness using youtube. *PLOS one*, *9*(10), e110171.

Noy, N. F., McGuinness, D. L. et al. (2001). *Ontology development 101: A guide to creating your first ontology.* Stanford knowledge systems laboratory technical report KSL-01-05 and ....

Ohlhorst, F. (2012). *Big data analytics: turning big data into big money* (Vol. 65). John Wiley & Sons.

Pathman, D. E., Crouse, B. J., Padilla, L. F., Horvath, T. V. & Nguyen, T. T. (2009). American recovery and reinvestment act and the expansion and streamlining of the national health service corps: a great opportunity for service-minded family physicians. *The Journal of the American Board of Family Medicine*, *22*(5), 582–584.

Pennington, J., Socher, R. & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543).

Qin, Y. & Zeng, Y. (2018). Research of clinical named entity recognition based on bi-lstm-crf. *Journal of Shanghai Jiaotong University (Science)*, *23*(3), 392–397.

Ren, Y., Teng, C., Li, F., Chen, B. & Ji, D. (2017). Relation classification via sequence features and bi-directional lstms. *Wuhan University Journal of Natural Sciences*, *22*(6), 489–497.

Seerden, X., Salmela, H. & Rutkowski, A.-F. (2018). Privacy governance and the gdpr: How are organizations taking action to comply with the new privacy regulations in europe? In *Ecmlg 2018 14th european conference on management, leadership and governance* (p. 371).

Souag, A., Salinesi, C. & Comyn-Wattiau, I. (2012). Ontologies for security requirements: A literature survey and classification. In *International conference on advanced information systems engineering* (pp. 61–69).

ul haq Dar, E. & Dorn, J. (2018). Ontology based classification system for online job offers. In *2018 international conference on computing, mathematics and engineering technologies (icomet)* (pp. 1–8).

Walczuch, R. M. & Steeghs, L. (2001). Implications of the new eu directive on data protection for multinational corporations. *Information Technology & People*, *14*(2), 142–162.

Wang, Y., Norcie, G., Komanduri, S., Acquisti, A., Leon, P. G. & Cranor, L. F. (2011). I regretted the minute i pressed share: A qualitative study of regrets on facebook. In *Proceedings of the seventh symposium on usable privacy and security* (p. 10).

Wen, T.-H., Gasic, M., Mrksic, N., Su, P.-H., Vandyke, D. & Young, S. (2015). Semantically conditioned lstm-based natural language generation for spoken dialogue systems. *arXiv preprint arXiv:1508.01745*.

Yaguinuma, C. A., Ferraz, V. R. T., Santos, M. T. P., Camargo, H. A. & Nogueira, T. M. (2010). A model for representing vague linguistic terms and fuzzy rules for classification in ontologies. In *Iceis (2)* (pp. 438–442).

Yang, X., Macdonald, C. & Ounis, I. (2018). Using word embeddings in twitter election classification. *Information Retrieval Journal*, *21*(2-3), 183–207.

Yepes, A. J. (2017). Word embeddings and recurrent neural networks based on long-short term memory nodes in supervised biomedical word sense disambiguation.

Journal of biomedical informatics, *73*, 137–147.

Zhao, Y., Dong, J. & Peng, T. (2009). Ontology classification for semantic-web-based software engineering. *IEEE Transactions on Services Computing*, *2*(4), 303–317.