A User-Centric and Sentiment Aware Privacy-Disclosure Detection Framework based on Multi-input Neural Network

A K M Nuhil Mehdy
akmnuhilmehdy@u.boisestate.edu
Boise State University
Boise, Idaho, USA

Hoda Mehrpouyan
hodamehrpouyan@boisestate.edu
Boise State University
Boise, Idaho, USA

ABSTRACT

Data and information privacy is a major concern of today's world. More specifically, users' digital privacy has become one of the most important issues to deal with, as advancements are being made in information sharing technology. An increasing number of users are sharing information through text messages, emails, and social media without proper awareness of privacy threats and their consequences. One approach to prevent the disclosure of private information is to identify them in a conversation and warn the dispatcher before the conveyance happens between the sender and the receiver. Another way of preventing information (sensitive) loss might be to analyze and sanitize a batch of offline documents when the data is already accumulated somewhere. However, automating the process of identifying user-centric privacy disclosure in textual data is challenging. This is because the natural language has an extremely rich form and structure with different levels of ambiguities. Therefore, we inquire after a potential framework that could bring this challenge within reach by precisely recognizing users' privacy disclosures in a piece of text by taking into account - the authorship and sentiment (tone) of the content alongside the linguistic features and techniques. The proposed framework is considered as the supporting plugin to help text classification systems more accurately identify text that might disclose the author's personal or private information.

CCS CONCEPTS

• **Security and privacy → Privacy protections**.

KEYWORDS

Privacy, Natural Language Processing, Neural Network

1 INTRODUCTION

Privacy is an ancient concept concerning human values that could be "intruded upon", "invaded", "violated", "breached", "lost", and "diminished"[29]. Each of these analogies reflects a conception of privacy that can be found in one or more standard models or theories of privacy. Users' privacy has been defined as "the right to be left alone" or being free from intrusion by the seclusion and non-intrusion theory[8, 32]. Even though privacy varies from individual to individual and each user may have different views of privacy, there is an imperfect societal consensus that certain information (e.g. personal information, situation, condition, circumstance, etc) is more private than the others (e.g. public statements, opinion, comments, etc)[4].

Recent advances in communication technologies such as messaging applications and social media [29] have resulted in privacy concerns [21] about analogous information amongst the users. In this era of digital communication, an increasing number of users are sharing information through text messages, emails, and social media without proper awareness of privacy threats and their consequences. Moreover, in the context of the information society, historical documents of entities (e.g. people, organization) are needed to be made public and shared among authorities every day [23]. In such cases, improper disclosure [1] of user's information could increase his/her security/privacy vulnerabilities, and the negative consequences of disclosing such information could be immense [7].

A recent data scandal involving Facebook and Cambridge Analytica reveals how personally identifiable information of up to 87 million Facebook users influenced voter's opinion [10, 25]. Likewise, millions of data breach incidents are reported all over the world and unfortunately most of them expose users' personal data [27]. Therefore, user-centric targeted attacks by exploiting the victim's Personally Identifiable Information (PII) has become a new kind of privacy threat in the present-day [31]. It's worth mentioning that United States is the number one destination for such user-centric targeted attacks based on recent statistics [28]. That being the case, users' data privacy has become one of the major concerns of today's world and the requirements for privacy measures to protect sensitive information about individuals have been researched extensively [3, 12–14, 19, 24].

As part of this efforts, researchers in the area of Natural Language Processing (NLP) have focused on developing techniques and methodologies to detect, classify, and sanitize private information in textual data. However, most of these works tend to solve these tasks by just detecting set of keywords, leveraging dictionaries of terms, or applying regular expression patterns. These types of detection do not consider the context and the relationship of the keywords in the text, therefore they result in high amount of false positive (e.g. a doctor's article about a disease is considered public and not private). However, it is considered sensitive and private when associated with other entities (e.g., a patient himself) in certain ways that yield different meaning and actually reveals someone's privacy. Therefore, its equally necessary to look into the keywords, data subject (i.e. users), authorship, tone, and overall

PrivateNLP 2020, Feb 7, 2020, Houston, Texas

[1] In this work, disclosure is defined as revealing personally identifiable information (e.g., name, address, age) or sensitive data (e.g., health, finance, and mental status) to others.

meaning of the content before classifying as privacy disclosure (Refer to Figure 1). While a few of the recent works are concerned with disclosure detection techniques by considering user-centric factors, most of them still omit other important decision-making factors such as sentiment and authorship of the content. Therefore, this paper aims to review the existing methodologies and techniques from the area of NLP and proposes a novel disclosure identification framework by keeping the following factors in mind:

- **Considering users-centric circumstances, tone, and authorship of content:** content having - sensitive information but no data subject, sensitive keywords but public ambience, analytical tone should not be classified as disclosure.
- **Checking sentence coherence and grammatical structure:** appearance of random keywords, ambiguous and meaningless information, or invalid utterances should not be classified as disclosure.

The rest of the paper is organized as follows: Section 2 contains the review on the related research works following some of their limitations we have observed. Section 3 describes about the dataset used in this paper. The methodology is described in detail in section 4. In addition, the detail of the deep neural network architecture, data cleaning, pre-processing, featurization, and the experiment is presented in section 5. Lastly, section 6 represents the experimental results following the conclusion.

I've been to two **clinics** and had my **pcp**. I've had an **ultrasound** only to be told it's a resolving **cyst** or a **hematoma**, but it's getting **larger** and starting to make my leg **ache**. The **PCP** said it can't be a **cyst** because it started out way too **big**. I am now **scared** and **afraid** of **cancer**...

Idaho is a state in the northwestern region of the United States. It borders the state of Montana to the east and northeast, Wyoming to the east, Nevada and Utah to the south, and Washington and Oregon to the west. To the north, it shares a small portion of the Canadian border with the province of British Columbia.

Don't be **scared** and do not assume anything **bad** as **cancer**. I have gone through several cases in my **clinic** and it seems familiar to me. As you mentioned it might be a **cyst** or a **hematoma** and it's getting **larger**, it must need some additional **diagnosis** such as **biopsy**...

Figure 1: Example of disclosure post, non-disclosure post, and highly similar to disclosure but actually a non-disclosure post (from top to bottom respectively).

2 BACKGROUND AND RELATED WORK

This section is a review on the state-of-the art research studies, related to information disclosure identification of individuals or organisations. Specially, the related literature which has been studied across different privacy domains such as finance, health, location, etc. We briefly described the research works which are related to detection, classification, and sanitization of private information in natural language text. We categorize the related works into three distinct groups based on their methodologies: i) Leveraging Dictionaries ii) Information Theory and Global Search iii) Machine Learning and Statistical Models.

The works under the category of dictionary utilization, leverage the linguistic resources such as privacy dictionary to automate the content analysis of privacy related information. Privacy dictionaries are used with existing automated content-analysis software such as LIWC [18]. Vasalou et al. proposes a technique that uses such a dictionary of individual words or phrases which are assigned to one or more privacy domains [30]. They showed that the dictionary categories could distinguish differences between documents of privacy discussions and general language by measuring unique linguistic patterns within privacy discussions (e.g., medical records, confidential business documents).

The researchers from the area of information theory utilize theories along with large corpus of words to automatically detect sensitive information from textual documents. [22] define sensitive information as the pieces of text that either reveals the identity of a private entity or refer to some confidential information of that entity. In their approach sensitive terms are those that provide more information than common terms due to their specificity. Therefore, the task is to quantify how much information each textual term provides, before identifying those as sensitive terms. Similar document sanitization tasks have been well addressed by Chakaravarthy et. al. where they represent a scheme that detects sensitive elements using a database of entities instead of patterns [5]. Each entity in this database (e.g., persons, products, diseases, etc.) is associated with a set of terms related to the that entity. Each set is considered as the context of an entity. For example, the context of a person type entity could be his/her birth date, name etc. Another research work by Abril et. al. that focuses on domain-independent unstructured documents has also been reviewed [1] where they propose to use a named entity recognition techniques to identify sensitive or private entities.

Detection of privacy leaks has also been well-addressed by machine learning and statistical techniques such as association rule mining [6]. In such an approach, (Chow et. al.) employs a model of inference detection using a customized web based corpus as reference where inferences are based on word co-occurrences. The model is then provided a topic (e.g. HIV - human immunodeficiency virus) and said to identify all the associated keywords. Hart et al. (2011) utilize machine learning techniques to classify full documents as either sensitive or non-sensitive by automatic text classification algorithms [9]. Their task is to develop an efficient and automated tool for enterprise data loss prevention (DLP) by keeping the sensitive documents secret. They introduce a novel training strategy called *supplement and adjust* to create an enterprise-level classifier

based on support vector machine (SVM) with a linear kernel, stop word elimination, and unigram methodology.

2.1 Our Contribution

The limitations of the current studies are based on the fact that they solely rely on the existence of keywords and neglect the sentence coherence, ignore grammatical validation, and disregard meaning inference in a piece of content. It has been addressed that these limitations, in some cases, result in miss classification and could be resolved by integrating parts-of-speech tags, dependency parse tree information, and word embedding[20]. However, a novel approach is required to take into the account the emotional tone or sentiment of the users that are hidden in the textual contents. For example, in Figure 1, the text from the red box is revealing someone's private (health) information (the patient has cancer) and the text from green box is about the Idaho state that represents some public ambiences. It's quite easy to distinguish these two piece of texts based on the keyword spotting techniques [2]. However, in another example, the text from the yellow box (comment from a doctor about cancer) has similar keywords as the patient's post, in the red box, containing valid word sequences and the presence of grammatical subjects (i.e. first person) with references etc. This piece of text is definitely not revealing private health situation (i.e. the doctor himself does not have cancer). Hence, it is quite challenging to distinguish between the types of contents without taking into the consideration the sentiment of the statements. To this end, this paper focuses on distinguishing highly similar contents based on the users' involvement, sentiment, authorship, and grammatical structure to classify texts containing someone's privacy disclosure. However, one of the assumptions of this work is: the proposed model does not solve all the privacy and security requirements of users by providing an entire threat model, rather it provides a better NLP tool to be integrated into any comprehensive privacy framework.

3 DATASET

We collected 10,000 users' (patients and doctors) posts from a public online health forum, based on the observation (inspired from the example of figure 1) that, patients' posts are somewhat disclosing their health status in that forum. Whereas, doctors' comments on patients' posts are highly similar content (having similar keywords and syntactic representation) but usually do not disclose doctors' health status (doctors' do not have those diseases). Therefore, we labeled patients' posts as disclosure (private) and doctor's comments as non-disclosure (public). For this paper, we crawled 5000 posts and 5000 comments and narrow down our privacy domain to health only. The length of the posts and comments varies from 10 words to more than 100 words comprised of several sentences.

4 METHODOLOGY

Combination of both linguistic operations and artificial neural network is the core of our methodology. A bigger picture of the framework is depicted in Figure 2. In this section, the data pre-processing, representation, and featurization steps are briefly explained following the detail of the neural network architecture.

4.1 Featurization and Data Representation

As can be seen from the examples in Figure 1 many domain specific keywords can be used in both private and public posts. This makes the problem particularly challenging because we cannot simply rely on the lexical items in the text; we have to consider the intent of the author of the text, and somehow determine if the intent was for the text to be public or private. To this end, we do custom tokenization and enrich our data with additional information using linguistic details such as syntactic dependency relations.

Tokenization. In many text-based natural language processing tasks, the text is pre-processed by removing punctuation and stop words, leaving only the lexical items. However, we found that the way people punctuate their texts helps give the clues as to whether or not it is a valid private or public information. Therefore, we use NLP Toolkit to tokenize the sentences in a customized way that ignores redundant tokens such as *",", ";–", "!!!", ":-)"* but keeps the important ones such as *",", ";", ":", ".", "he", "the", "in"* etc. This step of considering all the valid sequential tokens helps our model learn important arrangement of tokens for validating relationships of entities. This is somewhat in contrast to other text analysis literature where clearing off all the punctuation tends to improve task performance.

Syntactic Structure. In the experiments, dependency-parse-tree information is also utilized as additional underlying features that improved the performance of the neural network model. This helps the model to observe common sequence of tokens as well as co-occurrence of dependency tags. We use a Dependency Parser (DP) Toolkit to extract the syntactic relation information (which is different from, but in some ways similar to, entity relation information). This allowed us to enrich our data with dependency parse information.

Supplemental Features. In addition to the features mentioned above, more user specific features or meta data are prepared and provided to the extended variant of our models as supplimental input. Some of those auxiliary data are - i) number of pronouns ii) emotional tone iii) number of negations found in the post etc. This additional information are supposed to give the neural network model some distinguishable features about highly similar contents of different class.

4.2 Deep Neural Network Model

After doing all the necessary pre-processing steps the data is then fed into a multi-input deep neural network to learn the hidden patterns and features to distinguish between texts having disclosure and non-disclosure occurrences. It takes lexical (word tokens) features through one input, syntactical features (dependency parse tree information) through another input following a merging of those feature vectors. Later these vectors additionally get merged with supplemental (auxiliary) inputs before going through a further multi-layer perceptron stage. At the end of the deep neural network, a single neuron is used to provide the probability toward each of the above mentioned classes. More detail about the architecture is depicted in appendix C.

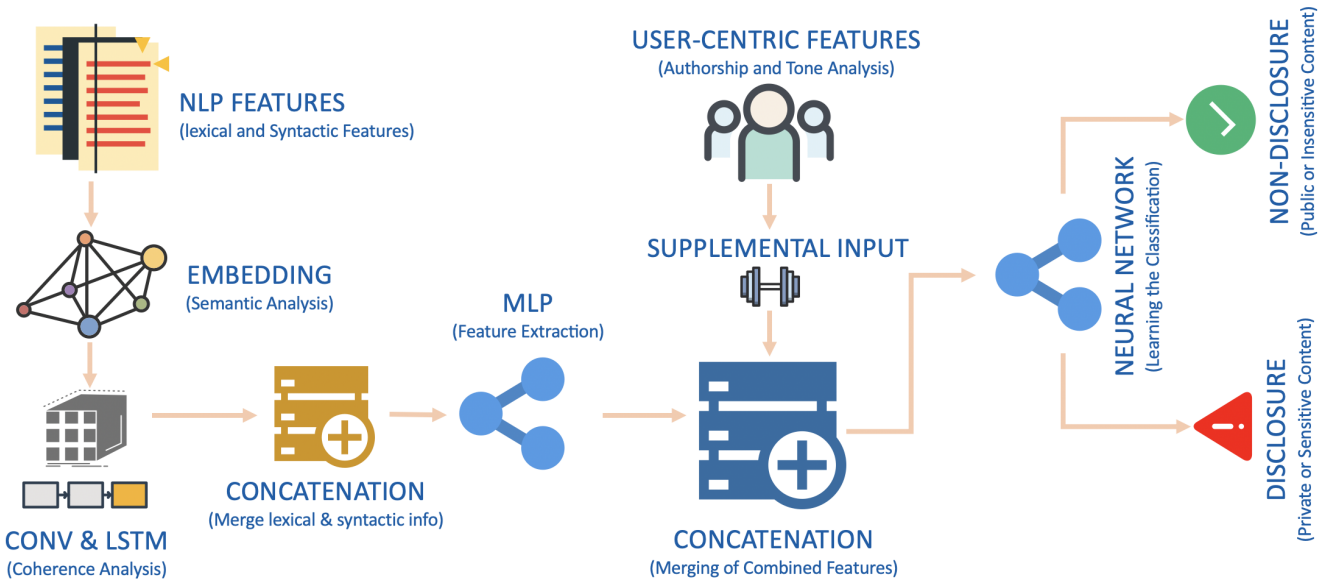


Figure 2: Bigger picture of the disclosure detection framework.

5 EXPERIMENT

In the data pre-processing step, we apply Spacy [26] to perform the linguistic operations on the text. The Keras functional API is utilized to create the multi-input architecture [17]. For implementing word embeddings, we use it's *Embedding* [16] layer where pre-trained word embedding (glove) is used with trainable flag set to true. In another input of the multi-input model, same type of embedding layer but without pre-trained vector, is used to learn the embedding space from the dependency parse tree information. For the *Convolution* on the information of the first input channel, we use the Conv1D layer [15] following a pooling layer just after it.

In the other input of the model, a long short term memory (LSTM) layer is used over the dependency parse tree information. The `concatenate` method of Keras then takes the output vectors from the convolution layer and the lstm layer and merges them into a single vector which then acts as the input to the fully connected layers. At this step, supplemental input, prepared by utilizing IBM Watson Tone Analyzer[11] are added with the concatenated vector following another stage of dense layers. Finally, a single neuron with sigmoid activation function outputs the probability of each class with 0.5 as the cutoff value. As false negatives of the classifier may bring dangerous consequences, it would be wise to lower this probability cutoff value towards the negative class, depending on the usage of the model. Detail of the hyperparameters are listed in appendix B.

6 RESULTS

Prior to experiment with the multi-input model, the classification task was examined using baseline models such as naive bayes classifier and simple convolutional neural network. Appendix A shows in detail the comparison of accuracy among all the models along with the model which uses user-specific supplemental input. The results show that, despite a lack of large amounts of labeled data, neural network based classifier can be trained that goes beyond simple keyword spotting and uses linguistic features to determine if a text contains a disclosure or not with an useful degree of accuracy. Moreover, it is observed that, integration of user-specific meta data to the models increases the classification accuracy, significantly (up to 97%). However, the generalizability of the model has not been well evaluated because of the lack of data set with similar characteristics (i.e., indistinguishable utterances yet carrying different meaning).

7 CONCLUSION

A practical model of privacy disclosure detection is in dire need by users in this era of social networks that results in activities such as online forum posting, emailing, text messaging etc. Accordingly, the development of algorithm and tools that helps identifying privacy disclosure in textual data is important. While many of these works in this area mainly focus on classifying textual data as public or private at the document level by just spotting keywords, only few of those are concerned with the the privacy detection, taking the users context into account.

ACKNOWLEDGMENTS

The authors would like to thank National Science Foundation for its support through the Computer and Information Science and Engineering (CISE) program and Research Initiation Initiative(CRII) grant number 1657774 of the Secure and Trustworthy Cyberspace (SaTC) program: A System for Privacy Management in Ubiquitous Environments

REFERENCES

[1] Daniel Abril, Guillermo Navarro-Arribas, and Vicenç Torra. 2011. On the declassification of confidential documents. In *International Conference on Modeling Decisions for Artificial Intelligence*. Springer, 235–246.

[2] Roy F Baumeister and Kenneth J Cairns. 1992. Repression and self-presentation: When audiences interfere with self-deceptive strategies. *Journal of Personality and Social Psychology* 62, 5 (1992), 851.
[3] Tom Buchanan, Carina Paine, Adam N Joinson, and Ulf-Dietrich Reips. 2007. Development of measures of online privacy concern and protection for use on the Internet. *Journal of the Association for Information Science and Technology* 58, 2 (2007), 157–165.
[4] Aylin Caliskan Islam, Jonathan Walsh, and Rachel Greenstadt. 2014. Privacy detective: Detecting private information and collective privacy behavior in a large social network. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society*. ACM, 35–46.
[5] Venkatesan T Chakaravarthy, Himanshu Gupta, Prasan Roy, and Mukesh K Mohania. 2008. Efficient techniques for document sanitization. In *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM, 843–852.
[6] Richard Chow, Philippe Golle, and Jessica Staddon. 2008. Detecting privacy leaks using corpus-based association rules. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 893–901.
[7] Emily Christofides, Amy Muise, and Serge Desmarais. 2009. Information disclosure and control on Facebook: Are they two sides of the same coin or two different processes? *Cyberpsychology & behavior* 12, 3 (2009), 341–345.
[8] Jamal Greene. 2009. The so-called right to privacy. *UC Davis L. Rev.* 43 (2009), 715.
[9] Michael Hart, Pratyusa Manadhata, and Rob Johnson. 2011. Text classification for data loss prevention. In *International Symposium on Privacy Enhancing Technologies Symposium*. Springer, 18–37.
[10] Alex Hern. 2018. Far more than 87m Facebook users had data compromised, MPs told. https://www.theguardian.com/uk-news/2018/apr/17/facebook-users-data-compromised-far-more-than-87m-mps-told/-cambridge-analytica. [Online; accessed 01-April-2019].
[11] IBM. 2019. IBM Watson - Tone Analyzer. https://www.ibm.com/watson/services/tone-analyzer/. [Online; accessed 01-December-2019].
[12] Adam N Joinson, Ulf-Dietrich Reips, Tom Buchanan, and Carina B Paine Schofield. 2010. Privacy, trust, and self-disclosure online. *Human–Computer Interaction* 25, 1 (2010), 1–24.
[13] Rezvan Joshaghani, Stacy Black, Elena Sherman, and Hoda Mehrpouyan. 2019. Formal specification and verification of user-centric privacy policies for ubiquitous systems. In *Proceedings of the 23rd International Database Applications & Engineering Symposium*. 1–10.
[14] Rezvan Joshaghani and Hoda Mehrpouyan. 2017. A model-checking approach for enforcing purpose-based privacy policies. In *2017 IEEE Symposium on Privacy-Aware Computing (PAC)*. IEEE, 178–179.
[15] Keras. 2018. Convolutional Layres - Keras Documentation. https://keras.io/layers/convolutional/. [Online; accessed 01-February-2019].
[16] Keras. 2018. Embedding Layres - Keras Documentation. https://keras.io/layers/embeddings/. [Online; accessed 01-February-2019].
[17] Keras. 2018. Guide to the Functional API - Keras Documentation. https://keras.io/getting-started/functional-api-guide/. [Online; accessed 01-February-2019].
[18] LIWC. 2018. Linguistic Inquiry and Word Count. https://liwc.wpengine.com/. [Online; accessed 01-February-2019].
[19] Naresh K Malhotra, Sung S Kim, and James Agarwal. 2004. Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information systems research* 15, 4 (2004), 336–355.
[20] Nuhil Mehdy, Casey Kennington, and Hoda Mehrpouyan. 2019. Privacy Disclosures Detection in Natural-Language Text Through Linguistically-motivated Artificial Neural Network. In *2nd EAI International Conference on Security and Privacy in New Computing Environments*. EAI.
[21] Joseph Phelps, Glen Nowak, and Elizabeth Ferrell. 2000. Privacy concerns and consumer willingness to provide personal information. *Journal of Public Policy & Marketing* 19, 1 (2000), 27–41.
[22] David Sánchez, Montserrat Batet, and Alexandre Viejo. 2012. Detecting sensitive information from textual documents: an information-theoretic approach. In *International Conference on Modeling Decisions for Artificial Intelligence*. Springer, 173–184.
[23] Ashley Savage and Richard Hyde. 2014. Using freedom of information requests to facilitate research. *International Journal of Social Research Methodology* 17, 3 (2014), 303–317.
[24] Arnon Siegel. 1997. In Pursuit of Privacy: Laws, Ethics, and the Rise of Technology. *The Wilson Quarterly* 21, 4 (1997), 100.
[25] Olivia Solon. 2018. Facebook says Cambridge Analytica may have gained 37m more users' data. https://www.theguardian.com/technology/2018/apr/04/facebook-cambridge-analytica-user-data-latest-more-than-thought. [Online; accessed 01-April-2019].
[26] Spacy. 2018. Linguistic Features - Named Entities. https://spacy.io/usage/linguistic-features#section-named-entities. [Online; accessed 01-February-2019].
[27] Statista. 2019. Number of U.S. data breaches 2014-2018, by industry. https://www.statista.com/statistics/273550/data-breaches-recorded-in-the-united-states-by-number-of-breaches-and-records-exposed/. [Online; accessed 01-April-2019].
[28] Symantec. 2019. 10 cyber security facts and statistics for 2018. https://us.norton.com/internetsecurity-emerging-threats-10-facts-about-todays-cybersecurity-landscape-that-you-should-know.html. [Online; accessed 01-April-2019].
[29] Herman T Tavani. 2007. Philosophical theories of privacy: Implications for an adequate online privacy policy. *Metaphilosophy* 38, 1 (2007), 1–22.
[30] Asimina Vasalou, Alastair J Gill, Fadhila Mazanderani, Chrysanthi Papoutsi, and Adam Joinson. 2011. Privacy dictionary: A new resource for the automated content analysis of privacy. *Journal of the Association for Information Science and Technology* 62, 11 (2011), 2095–2105.
[31] Ding Wang, Zijian Zhang, Ping Wang, Jeff Yan, and Xinyi Huang. 2016. Targeted online password guessing: An underestimated threat. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. ACM, 1242–1254.
[32] Samuel Warren et al. 1890. Louis Brandeis. The Right to Privacy. *Harvard Law Review* 4, 5 (1890), 1.

A RESULTS IN DETAIL

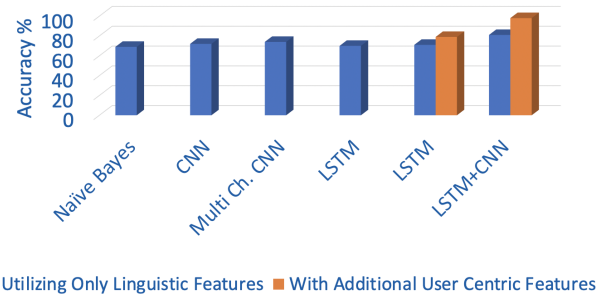


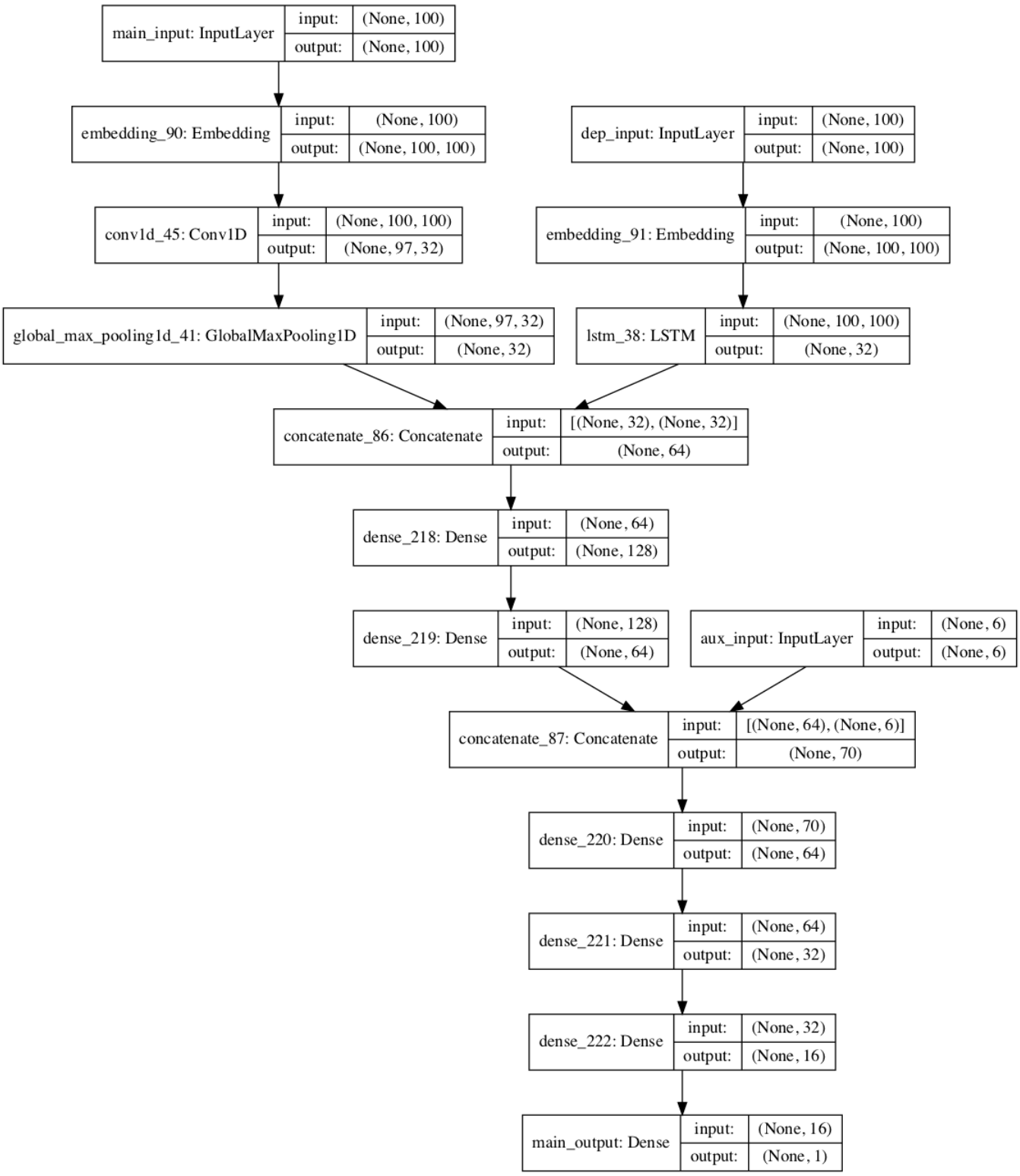
Figure 3: Accuracy of the model as a binary classification.

B MODEL HYPERPARAMETERS

Some hyperparamters worth mentioning are: pre-trained embedding with glove 100 dimensional embedding matrix having the capability of adjusting weights through the training iteration. Convolution with 32 filters with kernel size of 4. These layers have rectifier linear unit as activation function and followed by global max pooling technique. The LSTM layer contains 32 neurons with all the default settings as per the keras documentation. The first stage of dense layers after the first concatenation contains 128 and 64 neurons with rectifier linear unit as activation function. The second stage of dense layers contains 64, 32, and 16 neurons with same kind of activation function following a single output neuron with sigmoid as activation function. We train the model for 20 epochs providing the batch size of 32. The model also uses binary cross entropy as the loss function and rmsprop as the optimizer.

C NEURAL NETWORK ARCHITECTURE

Architecture of the Neural Network (automatically rendered by the Keras plotter) is given below.


main_input: InputLayer
input: (None, 100)
output: (None, 100)
embedding_90: Embedding
input: (None, 100)
output: (None, 100, 100)
dep_input: InputLayer
input: (None, 100)
output: (None, 100)
conv1d_45: Conv1D
input: (None, 100, 100)
output: (None, 97, 32)
embedding_91: Embedding
input: (None, 100)
output: (None, 100, 100)
global_max_pooling1d_41: GlobalMaxPooling1D
input: (None, 97, 32)
output: (None, 32)
lstm_38: LSTM
input: (None, 100, 100)
output: (None, 32)
concatenate_86: Concatenate
input: [(None, 32), (None, 32)]
output: (None, 64)
dense_218: Dense
input: (None, 64)
output: (None, 128)
dense_219: Dense
input: (None, 128)
output: (None, 64)
aux_input: InputLayer
input: (None, 6)
output: (None, 6)
concatenate_87: Concatenate
input: [(None, 64), (None, 6)]
output: (None, 70)
dense_220: Dense
input: (None, 70)
output: (None, 64)
dense_221: Dense
input: (None, 64)
output: (None, 32)
dense_222: Dense
input: (None, 32)
output: (None, 16)
main_output: Dense
input: (None, 16)
output: (None, 1)