Do My Emotions Influence What I Share? Analysing the Effects of Emotions on Privacy Leakage in Twitter

Manasi Mittal*, Muhammad Rizwan Asghar*, and Arvind Tripathi†
*School of Computer Science, The University of Auckland, New Zealand
†Information Systems and Operations Management, The University of Auckland, New Zealand

***Abstract*—Social media has become an integral part of modern-day society. With increasingly digital societies, individuals have become more familiar and comfortable in using Online Social Networks (OSNs) for just about every aspect of their lives. This higher level of comfort leads to users spilling their emotions on OSNs and eventually their private information. In this work, we aim to investigate the relationship between users' emotions and private information in their tweets. Our research question is whether users' emotions, expressed in their tweets, affect their likelihood to reveal their own private information (privacy leakage) in subsequent tweets. In contrast to existing survey-based approaches, we use an inductive, data-driven approach to answer our research question. We use state-of-the-art techniques to classify users' emotions, and privacy scoring and employ a new technique involving BERT for binary detection of sensitive data. We use two parallel classification frameworks: one that takes the user's emotional state into account and the other for the detection of sensitive data in tweets. Consecutively, we identify individual cases of correlation between the two. We bring the two classifiers together to interpret the changes in both factors over time during a conversation between individuals. Variations were found with respect to the kinds of private information revealed in different states. Our results show that being in negative emotional states, such as sadness, anger or fear, leads to higher privacy leakage than otherwise.**

***Index Terms*—Privacy, Emotions, Twitter**

I. Introduction

Social media sites are currently one of the most popular means of communication. There are numerous platforms, including Instagram, Facebook, Twitter and Reddit, that allow individuals to connect, communicate and collaborate with others for a variety of reasons. Social media serves as a way of connecting with friends and family, discussing political opinions or religious belief, and publishing regular updates about one's daily experiences. It is estimated that there are approximately 2.96 billion active Online Social Network (OSN) users currently[1]. OSNs play a vital role in the broadcasting of events and serve as an online forum for personal conversations, group discussions, and debates. These discussions and debates, which can draw a large number of users with different beliefs and philosophies, can lead to emotionally charged arguments. While many users understand their responsibility of safeguarding their own private data, we observe privacy leakage in these emotionally heated debates and discussions on social media platforms.

Twitter[2] is a micro-blogging social platform that presents an interesting case for the study of privacy breach as the majority of tweets are publicly accessible. Twitter boasts 330 million monthly active users as of quarter 1, 2019[3]. With so many users voicing themselves, there are many instances where data meant to be kept private, becomes public. Our work aims at probing into one of the major causes of such revelation, *i.e.,* users' emotional states. The transitions in terms of emotions and privacy leaks are also gauged over time within the context of conversations.

Privacy on social media platforms has been a concern for individuals, scholars, and practitioners. It has been noted that most users are unaware of the default privacy settings [1] and even oblivious to the disclosure of sensitive information [2]. Also, those who are privacy-aware share more content more often, perhaps being conscious of already having taken care of their privacy. Further, conflicting privacy settings and behaviours among OSN users generally result in the unwanted revelation of some users' private data [3]. Secondary leaks are a primary source of information leaks in such networks. One example is connections of the user who hold a public account posting about medical issues the user is going through or revealing date of birth through birthday wishes. In this way, family, friends, and connections of the primary user often reveal a lot about the user [3]–[5]. The social footprints are available in three types of social media data: users' profile attributes, their social context and ties, and their published content [1]. When it comes to social ties, the users' network has a lot to say about the users themselves. The principle of 'homophily' [6] is quite evident throughout and it is seen that people with similar privacy preferences come together in a network [7].

In general, user's activity can provide insights on certain characteristics of the user, such as political leaning [8]. The relationship between the personality traits of Twitter users and their role on the platform (listeners, popular, highly-read, and influential) was studied by Quercia *et al.* [9]. They showed an accurate way of predicting one's real-life personality only by

[1]https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users

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

[3]https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users

using three counts on their virtual profiles. Similar work was done by Sumner *et al.* [10] wherein the 'dark' or anti-social traits of Twitter users were predicted by monitoring the user's activity and applying linguistic inquiry.

In Cancerous tweets [4], Anderson *et al.* throw light on one particular type of privacy leak: sensitive medical information. Retweet ratio, geolocation and, tweet intent were used as metrics. Surprisingly, the majority of tweets relating to illness were shared by family members and friends of the patient without their consent. Chung *et al.* discussed privacy leakage in Event-Based Social Networks (EBSNs) [11]. They found that sensitive information like LGBT status of the users was clearly evident through the groups they were affiliated with. Thomas *et al.* examined Facebook for revelation of sensitive information [3]. They developed a classification system that used the information from conversations between friends and publicly displayed connections. It was devised to make predictions about the user's gender, age, political, and religious views, and media interests.

The work of Dong *et al.* [12] identifies behavioural analogy to psychological variables that are known to affect users' disclosure behaviour. It includes the sensitivity of the requested/shared information, the trustworthiness of the information audience, the appropriateness of the request/sharing activity, the sharing tendency of the receiver/information holder, as well as some contextual information. This is where our work comes into the picture. Specifically, we target one such variable that could potentially influence a user's privacy behaviour *i.e.,* their emotions.

A study conducted by Wang *et al.* [13] about regrets on Facebook found that many users posted regrettable content on Facebook. This included information about drug or alcohol usage, sensitive topics like religion and politics, use of profanity, family and personal matters, comments about employment, revealing secrets and lies. Such regrettable content can pose great risks to the user. It can lead to one losing one's job or even being robbed. The work also discusses posting regrettable messages under certain emotional states. We can validate the same using real data. There are multiple works that explore emotion as a form of disclosure [14]–[16] rather than being a potential cause for the same. This is where our work comes in and we investigate the cause and affect.

The rest of the paper is organised as follows. In Section II, we explain our data collection process. In Section III, we discuss existing approaches for emotion recognition and privacy leakage identification. Further, we describe our proposed methodology. In Section IV, we show our results that investigate our hypothesis surrounding emotions and privacy leaks. In Section V, we conclude the paper by summarising our work and findings as well as we provide research directions for future work.

II. Data Collection

There are essentially two elements we wish to correlate through our analysis: sentiment and privacy. Previous works have utilised crawling and scraping techniques for the purpose of collecting tweets. There are some ethical practices to be followed. As far as the two components are concerned, there have been numerous explorations of tweets and sentiments. For our preliminary analysis, we decided to go forward with publicly available datasets.

The datasets provided within the WASSA EmoInt shared task [17] were used. They consist of 4 text files corresponding to tweets belonging to 4 emotional categories, namely anger, joy, sadness, and fear.

The second dataset was created and published in public domain by Crowdflower[4]. It is much more diverse in terms of the number of categories of emotions. There are 13 possible labels for tagging tweets: anger, boredom, empty, enthusiasm, fun, happiness, hate, love, neutral, relief, sadness, surprise, and worry. The data is in the form of a CSV file with 40,000 rows and 4 columns including tweet ID, sentiment, author, and content.

We approached manual annotation for the assignment of privacy labels. Since the Crowdflower dataset is comparatively larger, we selected the first 10% (4000 tweets) of the dataset for privacy labelling. The rules for assigning the tags were in accordance with the table provided by Caliskan Islam *et al.* [14]. The authors had asked the AMT workers to label tweets as being private or non-private based on whether they belong to one of the given privacy categories or not.

III. Our Approach

A. Emotion Recognition: Binary Classes

We employ binary classification of emotion or sentiment. In our sentiment analysis, we have the categories 'positive' and 'negative'. The classes we looked at were pertaining to emotion, namely 'happiness' and 'sadness'. The WASSA dataset is very small to advance with our binary classification job. We settled on taking the Crowdflower dataset and dropping all rows belonging to other classes apart from 'happiness' and 'sadness'. We were left with a dataset containing 10,374 rows associated with the two class labels. We encoded the classes with '0' denoting 'sadness' and '1' denoting 'happiness'. The resulting dataset is very well-balanced with about 5,000 tweets in each class. We need to represent the text columns in a numerical manner for the model to understand. In our methodology, count vectors are defined as:

p = Sum of all feature count vectors with label 1

```
p = tf_train[y_train==1].sum(0) + 1
```

q = Sum of all feature count vectors with label 0

```
q = tf_train[y_train==0].sum(0) + 1
```

Log-count Ratio

```
r = np.log((p/p.sum()) / (q/q.sum()))
```

b: The ratio of positive and negative training cases

```
b = np.log(len(p) / len(q))
```

[4]https://data.world/crowdflower/sentiment-analysis-in-text

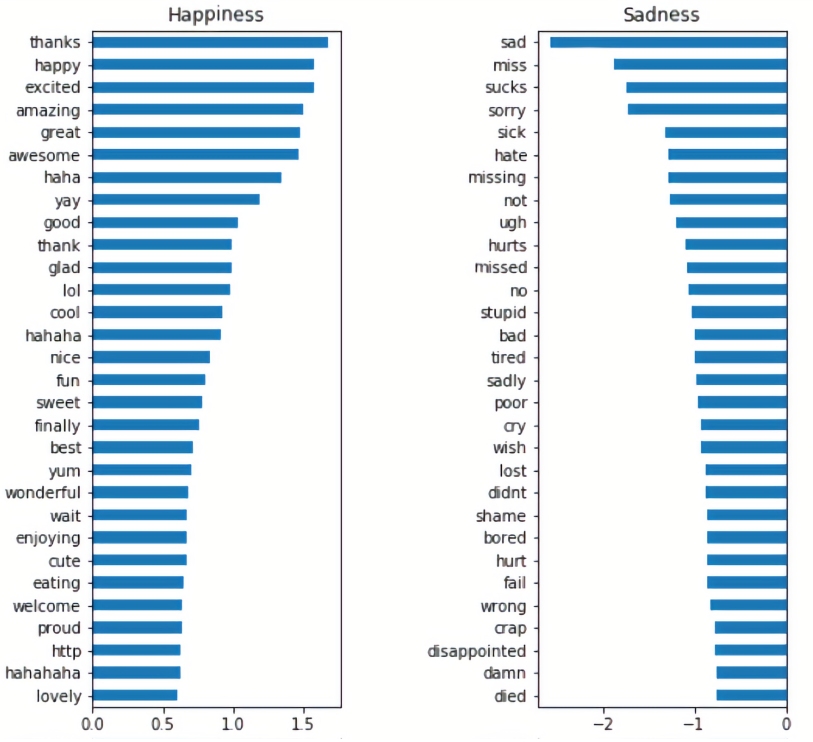

Fig. 1. Top 30 tokens relevant to each of the two classes for classification.

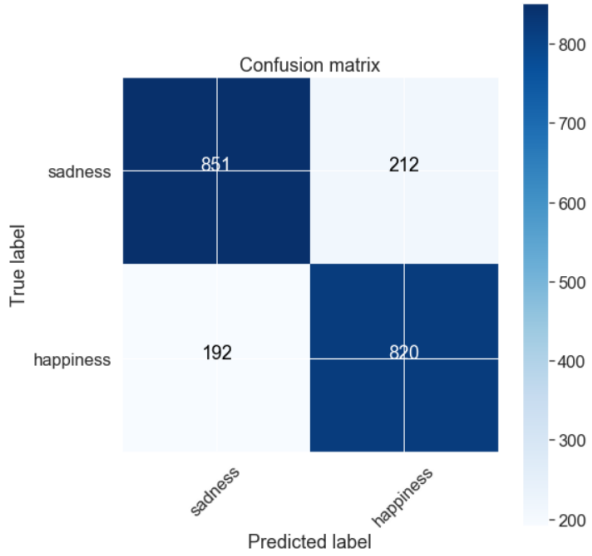

Fig. 2. Confusion matrix results for binary classification of emotions.

Now that we have all the necessary coefficients, we can figure out the predictions on the test set. We achieve an accuracy of **80.62%**. Instead of applying the equations to calculate the values of coefficients 'r' and 'b', we can train a model to learn and apply them. By fitting the model through logistic regression, we obtain an accuracy of **82.61%**. Although it might seem attractive to exploit complicated deep learning models like CNN and the likes for the goal of classification, as observed through our implementation, such models severely under-perform. The reason for this may be the data-hungry nature of such models that need huge amounts of data to deliver optimum results. We trialled a CNN model for the task of binary classification. The Keras neural network library was used for its implementation. We specified 80% of data to be used for training and the remaining 20% to be used for validation while training. The validation accuracy produced using verbose for each epoch peaked at 80% and continued to reduce thereafter. A confusion matrix gives us an overview of the correctly-classified and miss-matched data points post the application of classifier algorithm. We survey the matrix for the classifier given in Fig. 2.

We move on to analyse the most relevant tokens that the model is using for classification into the two categories. To this end, the logistic regression coefficients are mapped with their corresponding tokens. Consequently, they are sorted by importance. Fig. 1 illustrates the top 30 tokens used by the classifier for each of the categories. While most of them are relevant to their respective classes, the token 'http' seems out of place.

Surprisingly, removing the term 'http' from the vocabulary increases our accuracy by an entire percentage. The final accuracy of **83.6%** is sufficient enough to classify the Tweets as belonging to the 'sad' or 'happy' divisions.

B. Emotion Recognition: Multiple Classes

Although binary classification makes the tedious task of sentiment or emotion analysis much easier and accurate, it would

help to gain better judgement if we look at various emotions and their influence on privacy-compromising behaviour of individuals. Emotion can be a very subjective and ambiguous matter to identify through computational techniques.

Due to the highly unbalanced nature of the Crowdflower dataset (containing 13 labelled classes of emotion), we chose to go ahead with the WASSA dataset (containing 4 labelled classes of emotion) for our multi-class emotion prediction task. Before proceeding to the application of the models, we need to comprehend the data we have at hand.

It has 4 classes of data, each containing around a 1000 tweets. As the dataset is not highly imbalanced, over-sampling or under-sampling techniques should not be required before training.

Another variable of interest could be the tweet length. We targeted on identifying whether people Tweet longer or shorter texts when experiencing any of these 4 emotions.

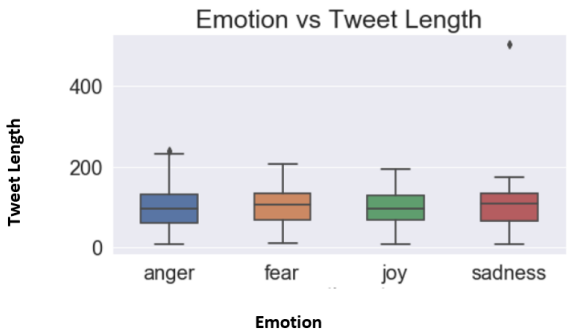

Fig. 3. Length of tweets across different classes in the WASSA dataset.

As we find in Fig. 3, there is no evident discrepancy in terms of the length distribution across categories. The tweets of the 'sad' class appear to be just a bit longer, but owing to the almost equal assortment among each class it does not count for a valuable feature of the texts.

Now the raw data was transformed into usable features that can serve as inputs and considerably improve the performance of models on unseen data. The TF-IDF score was used to portray the relative importance of a word in a particular text and the entire corpus. After feature vectorisation, there are 1281 features for each of the 3960 tweets. These features depict the TF-IDF scores for various uni-grams and bi-grams.

The Chi-square test is commonly used in research for testing relationships between categorical variables. The null hypothesis of the Chi-square test is that no relationship exists on the categorical variables in the data. By using this test, we find which terms are majorly associated with each of the sentiments. The resulting n-grams, as seen in Fig. 4, appear to be quite relevant to the distinct classes of emotion.

After we have applied all the aforementioned data transformation techniques, our data is equipped with the required labels and features. It is ready to be trained. We experimented with the numerous classification models and evaluated their accuracy. It is necessary to establish which model should be continued with for our analysis.

```
# 'anger':
  . Most correlated unigrams:
. bitter
. anger
  . Most correlated bigrams:
. im offended
. best revenge
# 'fear':
  . Most correlated unigrams:
. awful
. nightmare
  . Most correlated bigrams:
. watch amazing
. amazing lively
# 'joy':
  . Most correlated unigrams:
. optimism
. lively
  . Most correlated bigrams:
. lively broadcast
. amazing lively
# 'sadness':
  . Most correlated unigrams:
. sober
. lost
  . Most correlated bigrams:
. doing good
. grow weary
```

Fig. 4. Most correlated uni-grams, bi-grams for each class of emotions.

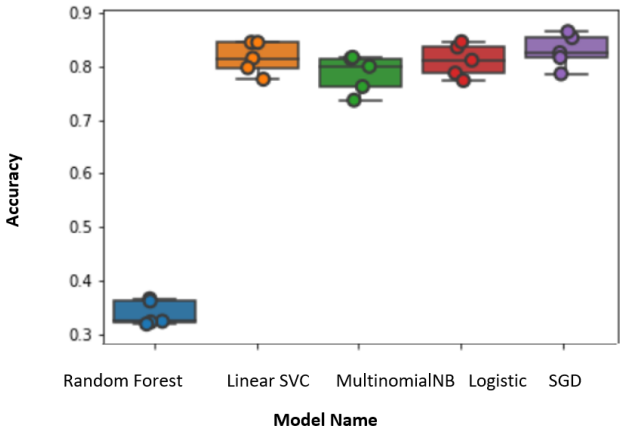

Fig. 5. Comparison of accuracy of various models for multi-class emotion tasks.

Fig. 5 clearly shows how the SGD classifier outperforms the other models. A comparison of the various classifiers according to mean accuracy, as shown in Table I, also gives us the same picture. With an average accuracy of around 83%, SGD does a pretty decent job of identifying the specific class of emotion a tweet would belong to.

Going forward with our best model, we create the confusion matrix in order to pinpoint the disparities between predicted classes and actual classes of the tweets. We observe that the vast majority of predictions end up along the diagonal of the matrix (Fig. 6) which is a good indicator of an efficient model.

C. Privacy Leakage Identification

Unlike the activity of sentiment or emotion classification, the detection of private versus non-private tweets using real

TABLE I
COMPARATIVE ANALYSIS OF PERFORMANCE OF VARIOUS CLASSIFIERS FOR MULTI-CLASS EMOTION DETECTION.

Model	Accuracy
Linear SVC	81.59
Logistic Regression	81.11
Multinomial Naive Bayes	78.63
Random Forest Classifier	33.84
SGD Classifier	82.95

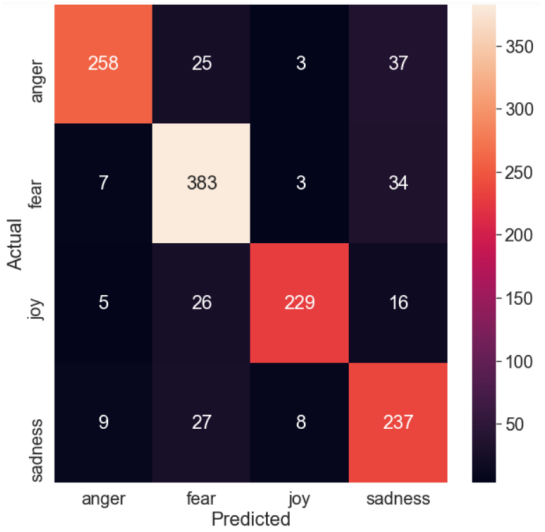

Fig. 6. Confusion matrix for multi-label emotion classification.

Twitter data is a rather unchartered one. Due to the lack of publicly available datasets, a portion of larger dataset (Crowdflower Dataset) was annotated manually (described in chapter-2). The tweets have been grouped into two categories 'private' and 'non-private' depending on whether or not they contain sensitive data. A more rigorous annotation scheme was followed later.

1) What makes a tweet 'private'?: Tweets may contain content like one's health condition, present emotional state, Personally Identifiable Information (PII) like address, date of birth, family, and personal details. Unaware of the potential threats that lie ahead, people end up disclosing such matter that then becomes publicly accessible. It is necessary to investigate the data and find which occurrences relate to such exposure.

2) Overview of Previous Privacy-scoring Approaches: There have been several works that aim at devising a scoring function or classifier to place sensitive tweets or information.

Many of them remain to be efficient in theory. They also have ready large annotated datasets. However, for our practical application that requires drawing a connection between emotions and privacy leaks, we must make use of a meticulous model. Our central focus is in anticipating if certain emotions lead to privacy leaks. We are not concerned with the amount or level of information revealed but rather with whether information is leaked or not. Therefore, for our fundamental analysis, we advance with binary classification. Since there is a slight dis-balance between the number of tweets labelled as private and non-private, we shall be considering metrics apart from accuracy for our final evaluation of model performance.

3) Binary Classifiers: Even the task of binary classification poses a great challenge as it is hard to distinguish what makes certain information sensitive or private and others non-sensitive. As with the preceding jobs of classification, we first made use of TF-IDF vectorisers and count vectorisers to transform the data into interpretable forms for the machine. The models performed better by making use count vector representation rather than TF-IDF vectors. While inspecting the data for linguistic, grammatical and textual features that may aid the performance of classifiers, the following observations were made:

1) **Punctuation:** We looked at the difference between tweets marked sensitive and those grouped as non-sensitive in terms of the punctuation used (question marks, exclamation points, period *etc.*) It was found that private tweets contain more exclamation points (5-8%) than non-private ones.
2) **POS-tagging:** We achieve this task with the help of NLTK's built-in functions and libraries. After applying Part-of-speech tagging to each corpus (non-sensitive and sensitive tweets), a very similar pattern in the distribution of tweets across the various parts of speech was discovered. Sensitive tweets in general contained more nouns and verbs than private ones. Tweets containing the noun 'http' are mainly non-private while private tweets contain the nouns 'work', 'time', 'headache', and 'friend'. In case of verbs, private Tweets had higher instances of 'get' and 'work' while non-private ones contain 'go' and 'do'.
3) **Uni-grams and Bi-grams:** Using the TF-IDF method, a word matrix is built. Then, logistic regression is used to rank the n-gram as per their importance. A higher score would indicate the feature is much more meaningful for our analysis. The tokens 'headache' are 'sick' are most correlated with the private category. A similar trend was seen in bi-grams and tri-grams where phrases like 'tummy hurts' and 'my head hurts' scored high. Also 'looking forward to' and 'im going to' indicated that people discussed a lot of their future activity as well.

Table II summarises the performance of various binary classification algorithms (applying count vectoriser and integrating relevant features).

TABLE II
SUMMARY OF PERFORMANCE: BINARY SENSITIVE TWEET CLASSIFICATION.

Model	F-1	Precision	Recall	Acc.
Logistic Regression	47.6	31.2	99.9	31.3
Multinomial Naive Bayes	82.1	79.9	84.5	73.4
SVM	82.8	80.8	84.8	74.5
Random Forest	84.9	77.2	94.2	75.7
XGBoost	85.2	75.5	97.7	75.4

4) Google's Bert Classifier: Since the task of classifying private content is a very non-standard one, we also probed into an unconventional, novel technique of classification. In the context of NLP language models, the Bidirectional Encoder Representations from Transformers (BERT) classifier has proven to achieve state-of-the-art results [18]. We apply the same to analyse content of twitter posts for private matters.

BERT is a method of pre-training language representations. The transformers bundle from Hugging Face will give us a pytorch interface for working with BERT.

The training set is divided up to use 90% for training and 10% for validation. Diverse NLP tasks can be tackled through the huggingface pytorch implementation as it contains interfaces specifically designed for each task. BertForSequenceClassification is used in our case. Along with the BERT model, it consists of an additional layer for tweet categorisation. Post loading the model, the parameters are set. They are assigned as follows:

Batch size: 32 Learning rate: 2e-5 Epochs: 4

We observe that with each epoch, the training loss is decreasing while the validation loss is increasing. This implies that the model is over-fitting on the training data. We can fix this by reducing the number of epochs we train our data on.

Performance We use Mathew's Correlation Coefficients (MCC) for evaluating performance. This scale indicated +1 as being perfect prediction and -1 being worst. We look into the MCC scores of predictions batch-wise as we can see in Fig. 7.

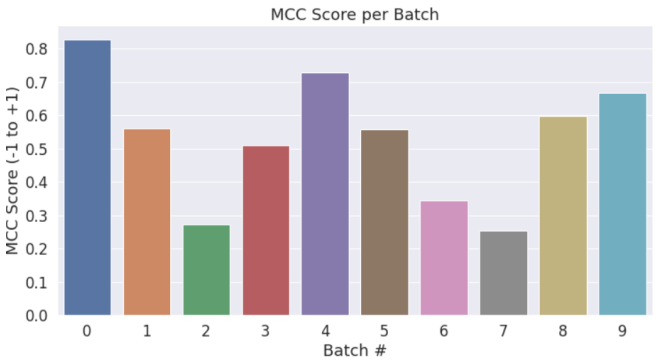

Fig. 7. BERT: Batch-wise MCC scores.

Then, we combine the above scores to produce the overall MCC score. It is 0.614, which is a pretty good measure on a range of -1 to +1. The F1 score obtained for predictions is an astounding 85.2%, which is actually better than that established through XGBoost.

5) Keyword-Specific Privacy Category Inquiry: It would be interesting to investigate the kinds of privacy leakage associated with the tweets. We can also correlate what kinds of emotions lead to what kinds of privacy leaks while interacting or conversing with others on Twitter. We adopted the keyword-based approach for identification of potentially sensitive tweets discussed by Wang *et al.* [19]. The categories include drug/alcohol, entertainment, family/personal, health/medical, obscenity, political, racism, relationship, religion, school life, sexual orientation, travel, and work.

IV. Our Results

By employing the models devised in the previous sections on the case of individuals tweets within our datasets, we examine if the emotional state of the users does in fact lead them to reveal sensitive information.

A. Varied kinds of Leaks across Varied Emotions

Here, we are primarily concerned with visualising what kinds of emotions are linked with what kinds of privacy divulges.

Applying the keyword-based privacy categorisation to our dataset gave us some intriguing results. We matched each word in each tweet with every keyword belonging to a particular file corresponding to one of the privacy categories. If the majority of the terms in the tweets were found in one file, the tweet was assigned the tag that identified with that file's privacy category.

The category to which maximum tweets belonged was work. We had removed the tweets that could not be classified within any of the privacy categories for further analysis. They had been labelled as 'no_tag_found' and could be easily identified and removed from the data frame. Around 50% of the tweets could not be recognised as belonging to any of the categories. Some compelling observations were made.

It was found that people when in a state of worry (Fig. 8) as well as sadness (Fig. 9) reveal more information related to work and medical matters.

It was also observed that in a state of happiness (Fig. 10), people revealed more information related to entertainment and family/personal matters.

These findings are very much in line with expected human behaviour. However, we believe that the scope of keywords alone in identifying privacy leakages is quite limited in nature. Subsequently, we see how applying the binary classifier for privacy leaks to the WASSA dataset (tweets labelled with anger, joy, sadness, and fear), we witness a higher correspondence between disclosure and certain emotional states.

B. Higher Incidence of Leaks in Certain Emotions

Here, we address the question "Does being in a particular emotional state while tweeting drive a user to reveal more information than otherwise?"

The emotional dataset used is the same that we used for multi-class allocation (Section II). We applied the BERT classifier to segment the tweets into private and non-private categories. Private class indicated that the tweet contained some form of content that could be considered risky on a public platform, either directly or indirectly compromising the user's data. Non-private indicated that the tweet could be considered safe on a public forum, it had little to no information about the user. The results of our analysis can be seen in Fig. 11.

It is clearly shown in Fig. 11 that when experiencing joy users tend to post "safe" tweets. They do not reveal much

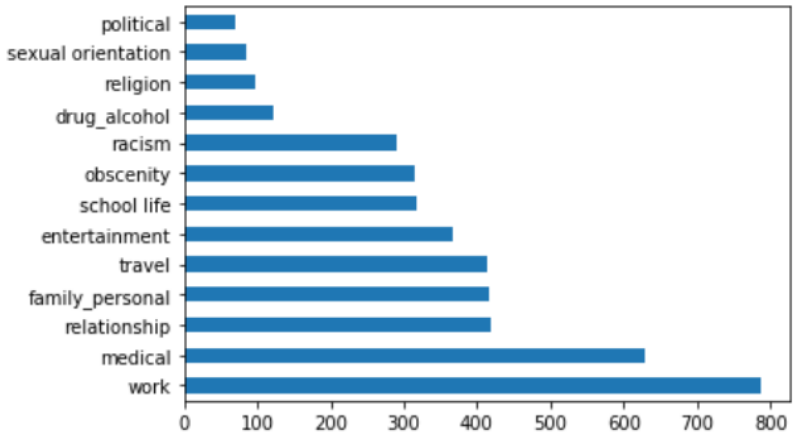

Fig. 8. Worry and privacy leaks.

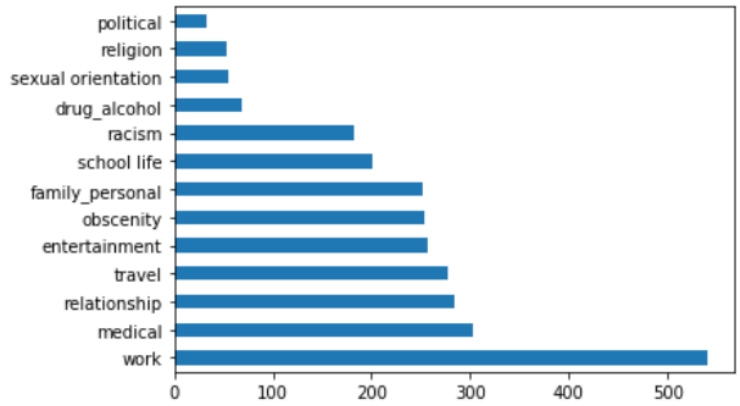

Fig. 9. Sadness and privacy leaks.

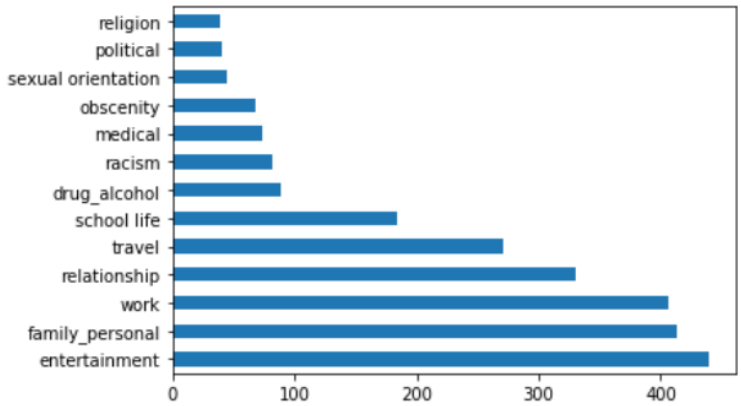

Fig. 10. Happiness and privacy leaks.

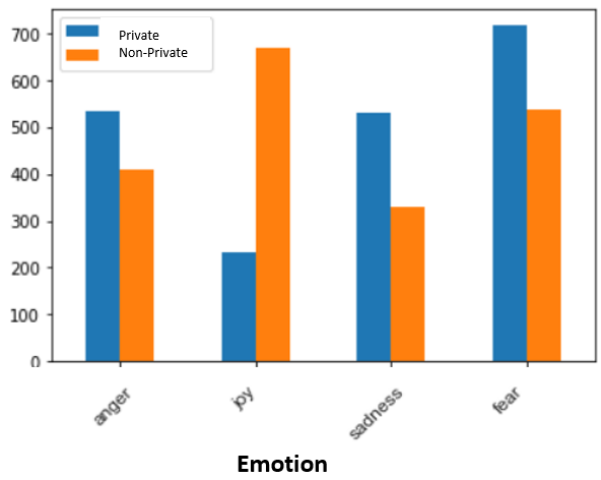

Fig. 11. Distribution of private and non-private tweets across emotions.

sensitive information. Whereas, the case of sadness presents quite the contrasting scenario. They tend to reveal more inside details. Although not to a significant extent, a similar observation is found in case of anger and fear.

It might be concluded through our data-based reasoning that emotions associated with negativity such as fear, anger, and sadness generally gravitate more instances of privacy leaks than their positive counterparts.

V. Conclusions and Future Work

Users on social media platforms are heterogeneous in their preferences for interactions with other users and revelation of private information. However, often users reveal private information not because of their preference but due to their emotional state. We noted that there have been instances of people getting fired and being robbed as a result of the same. The main objective of carrying out this inquiry was to tackle the root of such disclosures. This study shows significant correlations between the emotional states of users and privacy leaks. We also deduced that the occurrence of such exposure was higher in certain emotional states. Overall, this data-driven approach employed various approaches for emotion detection and identification of sensitive Tweets.

The research carried out serves as a way to many potential future directions. We wish to provide a remedy for this phenomenon so that individuals in certain emotional states are cautioned before tweeting or replying to another user. This could prevent the privacy of the user being compromised.

There are several other directions for future work in the space of privacy and OSNs. One such direction is to investigate the geographic distribution with respect to privacy behavior. People in different regions are found to be comfortable sharing different kinds of information. This could also mean a personalised recommendation of privacy settings and preferences according to cultural backgrounds.

We hope our work can be a stepping stone towards for identifying the elements necessary for maintaining a decent level of privacy on social media. It can also monitor emotional influences over conversing users and create a safe, risk-free space in social networks.

References

[1] T. Khazaei, L. Xiao, R. Mercer, and A. Khan, "Privacy behaviour and profile configuration in Twitter," in *Proceedings of the 25th International Conference Companion on World Wide Web*, 2016, pp. 575–580.

[2] R. Gross and A. Acquisti, "Information revelation and privacy in online social networks," in *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, 2005, pp. 71–80.

[3] K. Thomas, C. Grier, and D. M. Nicol, "unFriendly: Multi-party privacy risks in social networks," in *International Symposium on Privacy Enhancing Technologies Symposium*. Springer, 2010, pp. 236–252.

[4] M. D. Anderson, J. A. Adams, and E. R. Hooten, "Cancerous tweets: Socially sharing sensitive health information," in *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2014, pp. 648–653.

[5] M. Hirose, A. Utsumi, I. Echizen, and H. Yoshiura, "A private information detector for controlling circulation of private information through social networks," in *2012 Seventh International Conference on Availability, Reliability and Security*. IEEE, 2012, pp. 473–478.

[6] S. Guha and S. B. Wicker, "Do birds of a feather watch each other? homophily and social surveillance in location based social networks," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 2015, pp. 1010–1020.

[7] T. Khazaei, L. Xiao, R. E. Mercer, and A. Khan, "Understanding privacy dichotomy in Twitter," in *Proceedings of the 29th on Hypertext and Social Media*, 2018, pp. 156–164.

[8] H. Briola, G. Drosatos, G. Stamatelatos, S. Gyftopoulos, and P. S. Efraimidis, "Privacy leakages about political beliefs through analysis of Twitter followers," in *Proceedings of the 22nd Pan-Hellenic Conference on Informatics*, 2018, pp. 16–21.

[9] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft, "Our Twitter profiles, our selves: Predicting personality with Twitter," in *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*. IEEE, 2011, pp. 180–185.

[10] C. Sumner, A. Byers, R. Boochever, and G. J. Park, "Predicting dark triad personality traits from Twitter usage and a linguistic analysis of Tweets," in *2012 11th international conference on machine learning and applications*, vol. 2. IEEE, 2012, pp. 386–393.

[11] T. Chung, J. Han, D. Choi, T. T. Kwon, J.-Y. Rha, and H. Kim, "Privacy leakage in event-based social networks: A meetup case study," *Proceedings of the ACM on Human-Computer Interaction*, vol. 1, no. CSCW, pp. 1–22, 2017.

[12] C. Dong, H. Jin, and B. P. Knijnenburg, "Predicting privacy behavior on online social networks," in *ICWSM*, 2015, pp. 91–100.

[13] Y. Wang, G. Norcie, S. Komanduri, A. Acquisti, P. G. Leon, and L. F. Cranor, ""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*, 2011, pp. 1–16.

[14] A. Caliskan Islam, J. Walsh, and R. Greenstadt, "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*, 2014, pp. 35–46.

[15] X. Song, X. Wang, L. Nie, X. He, Z. Chen, and W. Liu, "A personal privacy preserving framework: I let you know who can see what," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018, pp. 295–304.

[16] G. Canfora, A. Di Sorbo, E. Emanuele, S. Forootani, and C. A. Visaggio, "A NLP-based solution to prevent from privacy leaks in social network posts," in *Proceedings of the 13th International Conference on Availability, Reliability and Security*, 2018, pp. 1–6.

[17] S. M. Mohammad and F. Bravo-Marquez, "WASSA-2017 shared task on emotion intensity," *arXiv preprint arXiv:1708.03700*, 2017.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[19] Q. Wang, H. Xue, F. Li, D. Lee, and B. Luo, "# DontTweetThis: Scoring private information in social networks," *Proceedings on Privacy Enhancing Technologies*, vol. 2019, no. 4, pp. 72–92, 2019.