Lindrit Kqiku, Marvin Kühn, Delphine Reinhardt. From Sentiment to Sensitivity: The Role of Emotions on Privacy Exposure in Twitter. In: Proceedings of the 2nd ACM Workshop on Open Challenges in Online Social Networks (OASIS, HT workshop), pp. 10–15, 2022.

From Sentiment to Sensitivity: The Role of Emotions on Privacy Exposure in Twitter*

Lindrit Kqiku
Institute of Computer Science, University of Göttingen
Göttingen, Germany
Campus Institute Data Science (CIDAS)
Göttingen, Germany
kqiku@cs.uni-goettingen.de

Marvin Kühn
Institute of Computer Science, University of Göttingen
Göttingen, Germany
marvin.kuehn@stud.uni-goettingen.de

Delphine Reinhardt
Institute of Computer Science, University of Göttingen
Göttingen, Germany
Campus Institute Data Science (CIDAS)
Göttingen, Germany
reinhardt@cs.uni-goettingen.de

Abstract

Online Social Networks (OSNs) are a vital part of users' daily lives. Users share content in OSNs increasingly more and in various emotional states. In this work, we explore the role of emotions on privacy exposure and we integrate it as an additional learning parameter in tweet sensitivity recognition. To this end, we first use BERT based classification techniques to recognize six basic emotions in tweets. Using our trained sentiment model, we further perform sentiment inference on a sensitivity dataset and integrate the sentiment in the BERT classification model to classify the tweets according to their sensitivity. We then compare the standard sensitivity recognition models' results (with their tweets only) against the extended model that integrates the sentiment features in sensitivity recognition. We demonstrate that by including sentiment features in sensitivity analysis, our approach leads to about a 3% increase of f-1 score in contrast to using our base sensitivity classification, i.e., from 83.96% to 87.01% f-1 score. We further demonstrate a correlation between *anger* and *disgust* emotions with *sensitive* tweets, as well as, *joy* and *surprise* with *non-sensitive* tweets.

CCS Concepts: • **Security and privacy → Privacy protections**; • **Computing methodologies → Neural networks**; **Supervised learning by classification**; **Natural language processing**.

*The final publication is available at ACM via https://doi.org/10.1145/3524010.3539501

OASIS'22, June 28, 2022, Barcelona, Spain

ACM ISBN 978-1-4503-9279-2/22/06…$15.00
https://doi.org/10.1145/3524010.3539501

Keywords: Privacy, sensitivity analysis, sentiment analysis, social networks, neural networks

ACM Reference Format:
Lindrit Kqiku, Marvin Kühn, and Delphine Reinhardt. 2022. From Sentiment to Sensitivity: The Role of Emotions on Privacy Exposure in Twitter. In *Open Challenges in Online Social Networks (OASIS'22), June 28, 2022, Barcelona, Spain.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3524010.3539501

1 Introduction

The widespread of *Online Social Networks* (OSNs), causes difficulties for users to control their own data while still keep using them. Twitter, in particular, is used by many individuals to share content online. Moreover, the number of monthly active users in Twitter is expected to keep increasing in the following years [7]. A considerable amount of shared content is sensitive and may lead to users' privacy violations. Moreover, users tend to often regret about their publicly shared tweets [20, 28]. Just deleting regrettable tweets is not enough and may backfire. They can easily be marked by malicious parties considering that such tweets have a higher chance to be damaging to the users. User's sensitive information can be at risk also from within their followers or friend lists. Therefore, it is of utmost importance to automate the recognition of sensitive content about to be posted in OSNs. As shown in [1], several works concluded that negative [5, 20, 28, 31] and positive emotions [5, 20] belong to some of the most common types of information that lead to regrets in OSNs. Sensitiveness modeling and analysis is particularly challenging due to the highly individual users' perception of privacy [9, 12], complexity of short and diverse texts, as well as, several other complex factors, e.g., different users' privacy concerns and attitudes, culture, etc.

In this paper, we aim to assess the sensitivity of text content with the inclusion of sentiment features. Two more recent works have been proposed to automate the recognition of sensitive tweets [16, 27]. Mittal et al. [16] investigated the role of four basic emotions on privacy exposure. They [16], however, did not incorporate the sentiment features as

a learning parameter in sensitivity recognition. Wang et al. [27] combined the sentiment features derived from Stanford sentiment tool and the tweet itself to determine tweets' sensitivity. Besides considering two more emotional states, our approach, in one hand, fills the gap of Mittal et al. [16] by integrating sentiment features, as a learning parameter, in a BERT classification model, to measure sentiment's added value in sensitivity recognition. In the other hand, by using more recent techniques than Wang et al [27] approach, we also are able to determine the possible advantage that BERT classification model could have as opposed to the GloVe and LSTM classification model of Wang et al. in a dataset annotated by users according to sensitivity of tweets [27]. Our study can be summarized as follows:

- We applied BERT classification model to train and evaluate the sensitivity model in recognising *sensitive* and *non-sensitive* tweets.
- We applied BERT classification model to train and evaluate the sentiment model in recognising sentiment of a tweet according to *anger, disgust, joy, surprise, fear*, and *sad* emotions.
- We then explored the correlation between *sensitive* and *non-sensitive* tweets with their emotions using the aforementioned BERT sensitivity and BERT sentiment classification models.
- We further estimate the inclusion of sentiment features derived by performing sentiment inference of our sentiment classification model in sensitivity recognition.

Our approach leads to a 87.01% f-1 score and 69.83% *Matthew's Correlation Ccoefficient* (MCC) score. We demonstrate that the inclusion of sentiment inference in sensitivity recognition increases the performance of sensitivity recognition by 3.05% in terms of f-1 score, respectively, by 2.11% under MCC score. We also show that users are more prone to share a *sensitive* tweet under *anger* and *disgust* emotions, and more *non-sensitive* tweets under *joyful* and *surprising* emotions. A tweet recognised as *fear* tweet can be almost evenly be a *sensitive* or *non-sensitive* tweet, whereas a tweet recognised as *sad* has a hardly noticeable tendency to be *sensitive*, however, a clear separation can not be made.

The rest of the paper is structured as follows. We dive into *Natural Language Processing* (NLP) background knowledge and some of the state-of-the-art NLP techniques in Sec. 2, related work in Sec. 3, before we introduce the used datasets in Sec. 4. We next outline our approach and used models and show our models' results in Sec. 5. We explore limitations and future directions in Sec. 6 and lastly summarize in Sec. 7.

2 Preliminaries

Multiple methods have been developed to solve NLP tasks over the years, especially with the advancement of deep learning. The earlier deep learning methods such as *Recurrent Neural Networks* (RNNs) [13, 22] and *Long Short-Term Memory* (LSTM) networks [11] computed the sequential input data in a sequential order. They made up for a long time as go-to approaches in NLP. In the recent years, *Graph Neural Networks* (GNNs) [21] and sequential transfer learning models [23] were introduced, which led in turn to a breakthrough with their performance in downstream tasks and their parallelization capabilities. Transformers models like *Bidirectional Encoder Representations Transformers* (BERT) [8], GPT-2 [18], XL-Net [30], and GPT-3 [4] moved the NLP field even further.

Due to its impact in the NLP field, BERT can be considered as a stepping stone model. BERT model [8] is based on the Vaswani et al. [23] approach. The model is pre-trained on a broad dataset and is subsequently fine-tuned via updating the gradients of pre-trained data or trained further before it proceeds with a language modeling task [8].

BERT, its light version DistilBERT [19], and thousands of other transformer based models can be applied using the Hugging Face library [29]. BERT comes in two main variants, namely, BERT_{BASE} and BERT_{LARGE}. The former uses 12 encoder blocks, 768, embedding dimensions, 12 attention heads, and 110 million parameters, whereas the latter has 24 encoder blocks, 1024 embedding dimensions, 16 attention heads, and 340 million parameters. They basic architecture however is the same. The authors of DistilBERT [19], in the contrary, aim not primarily for performance improvement but for reduced speed and size. DistilBERT has less parameters and six encoder blocks. It runs faster than BERT and retains over 97% of BERT's performance on language understanding tasks.

3 Related Work

Several works investigated the identification of privacy leakage information using *Machine Learning* (ML) in text content. Some of them focused on building topic-based classification models to predict whether tweets fit in a set of pre-defined sensitive topics [25, 26]. Several others used a more generic approach (1) to classify tweets in either private or public information [5, 16, 27], (2) to classify deleted tweets whether they are damaging or non-damaging tweets [15], (3) to classify vacation and drunk pre-selected tweets whether they are revealing sensitive content or not [14], or (4) to classify whether tweets are likely to be deleted or not [31].

Mao et al. [14] developed a model to identify potential private tweets using naive Bayes and SVM classifiers. Their built classifiers predict whether vacation and drunk driving tweets are sensitive or not. In addition, *Name Entity Recognition* (NER) tool was used to extract location name, person and time information. However, the models are trained for only three specific topics.

In another study [5], authors considered user-based tweets, topic matching extraction, NER tool, and sentiment analysis to classify users according to one of three labeled categories (i.e, first, second, and third privacy score) using AdaBoost and Naive Bayes ML classification models. They collected users' relationships and their activity on Twitter. Tweets were also analysed in terms of their metadata (i.e., hashtags and user references). Lastly, they randomly selected some users along with random selected tweets from their created database and asked the users to label their tweets according to a set of sensitive categories such as location, medical, drug/alcohol, emotion, etc. To this end, they determined where do users fit in terms of one of the three privacy scores, by examining how many tweets are sensitive from a user's total shared tweets.

Zhou et al. [31] predicted whether a tweet will be possibly deleted by analysing the features of deleted tweets and user preferences with regards to a set of ten pre-defined common sensitive topics (e.g., cursing, health, job, health) that were previously defined by another study [28]. A tweet was in turn classified as private if it contained terms from the sensitive topics. Naive Bayes was the best performing conventional ML model.

In two other studies [25, 26], authors considered further classifying private tweets into a set of either 13 or 14 pre-selected topic categories. They used *Bag-of-Words* (BOW) and *Term Frequency-Inverse Document Frequency* (TF-IDF) to extract the features and fit them to naive Bayes for classification. The tweets were classified binary to one of given categories. Besides ML techniques limitation, a further constraint of their approach was also the assumption that the private tweets were already pre-selected into one of the pre-defined categories.

Minaei et al. [15] considered content sensitivity protection from another point of view. Their domain of work was based on the observation that when users regret posting sensitive content, they delete the content. However, the act of deletion shows that the post is sensitive to the user and can be used by malicious parties. The authors thus proposed an approach to occasionally delete non-damaging posts in order to confuse the malicious parties about whether the tweets were actually sensitive or just a disguise. They used BERT to identify the damaging tweets.

More similar to our research line, certain studies examined content sensitivity according to a binary categorization as either private or public content [16, 27]. In a more recent study [27], authors used word embedding and RNN methods, specifically LSTM, to classify private tweets in either private or public categories. They further calculated a privacy score that included content sensitivity classification, sentiment analysis, and user's preferences. Mittal et al. [16] analysed particularly the role of users' sentiment in revealing sensitive content in tweets. They used BERT, as a more recent transfer learning technique to classify tweet sensitivity and the sentiment. Sentiment analysis over the WASSA-2017 dataset [17] was used to predict whether a tweet belongs to one of four sentiment categories (i.e., *text, fear, joy* and *sadness*). They however did not include the sentiment analysis in enhancing sensitivity classification and as compared in used a way smaller dataset than the one that we considered, as discussed in Sec. 4.1.

In summary, our approach goes to a similar line as Wang et al. [27] and Mittal et al. [16]. However, the focus of our paper is in fine-tuning transformer based models to enhance the classification of users' tweets in sensitive and non-sensitive classes, by coupling together sentiment in sensitivity prediction. Our approach is intended as a generic suggestion mechanism before a tweet is shared and neither on just a set of sensitive topic-based classifications (as in [25, 26]), nor on protection in the context of deleted tweets after they are already shared (as in [15]).

4 Datasets

In what follows, we explore two datasets that we used for sentiment and sensitivity recognition.

4.1 WASSA Dataset

We have chosen WASSA-2018 [6] dataset for sentiment analysis due to its compact size of labelled emotions and their vast amount of tweets per emotion. WASSA-2018 dataset consists of 191,731 tweets labeled according to six categories (*anger, sad, fear, joy, surprise, and disgust*). The original published dataset is split in training (i.e, 153,383 tweets), trial, and test data. We however used only the training set of the data (i.e., 80% of the data), since the trial and test labels were not available to us then. The dataset has a good distribution between the labels. The difference between the smallest and the largest set of labels is 18.75%. Using the Twitter API, the tweets were retrieved by filtering the emotion words or their corresponding synonyms along with some causality keywords, i.e, each of those tweets after the emotion trigger words has to contain either "that", or "because", or "when" keywords. This way, authors [6] assume that a tweet is likely to describe the cause of the emotion in a more implicit tone. Once the tweets were categorized using the aforementioned filtering method, the trigger emotion words were removed from the tweets. Thus, the models would supposedly learn to infer pattern implicitly from the context of the tweets. The former version of the dataset, named, WASSA-2017 [17] contained four categories with only 7,097 tweets distributed between *text, fear, joy* and *sadness* labels. However, due to its small sample size, smaller number of emotions, and non-implicit form, we used the WASSA-2018 in our assessments.

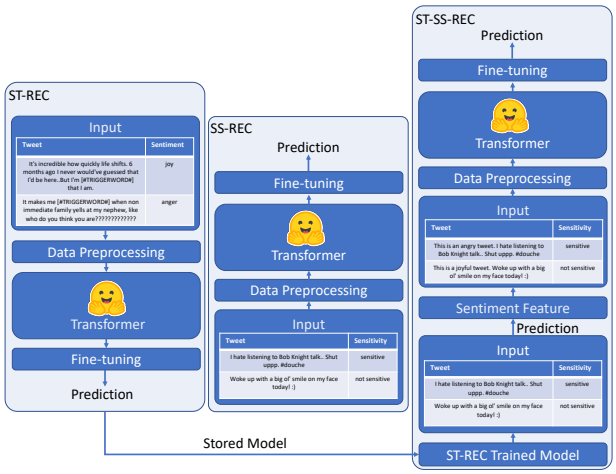


Figure 1. Pipeline of our solution. ST-REC is trained and evaluated on WASSA-2018 dataset, whereas SS-REC and ST-SS-REC on DTT dataset. ST-REC refers to the sentiment recognition, SS-REC refers to the sensitivity recognition, and ST-SS-REC refers to the inclusion of sentiment in sensitivity recognition.

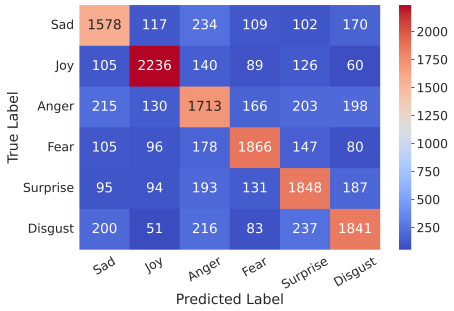


Figure 2. Confusion matrix in our sentiment recognition analysis (ST-REC) using WASSA-2018 dataset.

4.2 *#DontTweetThis* (DTT) Dataset

Wang et al. [27] conducted a user study in which users labeled tweets according to its sensitivity. Before asking participants for labeling of tweets, they pre-filtered the potentially private tweets. They generated an extended dictionary that is primarily based on established private topics from other sources [5, 14, 25, 31]. Each tweet was labeled separately from three different participants. Only the tweets that preserved a consistent consensus were taken into account. The dataset consists of 3,008 labeled tweets in *sensitive, maybe, non-sensitive* categories. 61 of them were labeled by users as *maybe*, 1,436 as *sensitive*, and 1,512 as *non-sensitive*. The undecidable (labeled as "maybe") tweets were omitted. We got from the authors, a version of the dataset that has an even distribution of tweets between *sensitive* and *nonsensitive* tweets, i.e., 1,435 tweets for each of the two categories. Hereafter, we use the terms *sensitive* interchangeably with *private* and *nonsensitive* with *public* as the authors of the dataset [27].

5 Design and Evaluation

We first investigate the models that recognize sentiment and sensitivity of tweets separately before we incorporate the former into the prediction of the latter, as illustrated in Fig. 1.

We used BERT transformer models to classify tweets according to their sentiments and their sensitivity. We used 80% of the dataset for training, 10% for validation, and the left 10% for testing. We used HuggingFace open-source library [29] for designing PyTorch based models. PyTorch is an open source ML framework, initially developed by Facebook's research team. We used the provided classifications of Hugging Face library, i.e., *BertForSequenceClassification* (BSC) to fine-tune them for sentiment and sensitivity classifications. BSC model consists of a BERT model as the first layer, a dropout layer that is used during training to improve regularization [10], and a linear classifier layer that applies a linear transformation to incoming tweets.

Authors of BERT [8] recommend the following values for fine-tuning models:

1. Batch size: 16, 32
2. (Adam) Learning rate: 5e-5, 3e-5, 2e-5
3. Number of epochs: two, three, four

We considered the aforementioned parameter ranges. We reached promising results with a batch size of 32, a learning rate of 2e-5, and four epochs. We further use those parameters in our recognition tasks.

In what follows, we outline in more details the designed and evaluated classifications task-wise.

5.1 *Sentiment Recognition* (ST-REC)

We trained and evaluated BSC classification model in WASSA-2018 dataset to detect the sentiment of tweets according to six categories. We reached the average weighted f-measure score of 72.27%.

In Fig. 2, we illustrate the summarized emotion predictions with their values across each emotional state. We observe that tweets labeled as *sad* might to a degree be confused with *anger* tweets more than the others and vice versa. Moreover, *disgust* labeled tweets may be occasionally misinterpreted as *surprise* or *anger* tweets more than others.

We finally saved the trained model, its configuration and tokenizer using *save_pretrained* method of Hugging Face library. The fine-tuned model and its corresponding saved vocabulary were in turn loaded in a later stage for sentiment recognition (Sec. 5.3). Both, the model itself, and its tokenizer, were loaded using *from_pretrained* method.

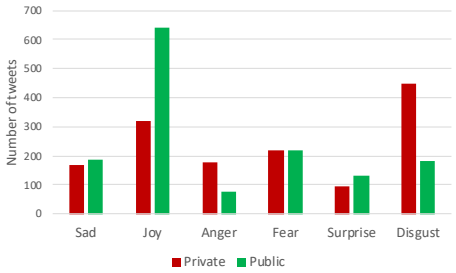


Figure 3. The correlation between public and private tweets and their emotions.

5.2 *Sensitivity Recognition* (SS-REC)

In this step, we fit the tweets into the developed models to train and evaluate BSC model. We investigated BERT_{BASE} and BERT_{LARGE} as the first layer of BERT classification model, by modifying only the first layer of BSC. We reached however better results with BERT_{BASE} transformer. Thus, we kept using the standard BSC designed model rather than our slightly modified model over the DTT dataset for a more extensive evaluation.

We evaluated our approach over five runs. For each run, we reshuffled the DTT dataset randomly using shuffle method of Python. Our model reached the f-1 score of 83.96% and MCC score of 67.72% over five run. Thus, we demonstrate that privacy exposure in Twitter can be determined fairly accurately by the tweet alone.

5.3 *Sentiment and Sensitivity Recognition* (ST-SS-REC)

The goal of this step is to further enhance the sensitivity recognition by including sentiment analysis. Before integrating the sentiment recognition learning parameter in sensitivity recognition, we first explore the occurrence of emotions in *sensitive* and *non-sensitive* tweets. To this end, we use the fine-tuned ST-REC model (Sec. 5.1) but this time over DTT dataset instead. In Fig. 3, we observe that users under certain emotions are more prone to leak sensitive information. Particularly, *anger* and *disgust* emotional connotations have a higher tendency to lie in the *sensitive* side. In contrast, *joyful* and *surprise* tweets lead to *non-sensitive* content. A distinction can not be made for *fear* and *sad* emotional states.

The distribution of emotions between *private* and *public* tweets was also investigated by Mittal et al. [16]. They explored four emotions, namely, *sad, joy, anger* and *fear*. They also found that *anger* tweets correlate more with privacy exposure and *joy* based tweets with being *public*. We diverge in the cases of *sad* and *fear* based tweets. In our study, *fear* and *sad* based tweets are almost evenly distributed between *private* and *public* tweets, in contrast to being more related to private information in their study.

We further integrate the sentiment features in sensitivity recognition approach as follows: We (1) use the fine-tuned sentiment classification model in WASSA-2018 (as introduced in Sec. 5.1), (2) we then perform sentiment inference on DTT dataset, and (3) lastly train and evaluate our sensitivity models in DTT datatset by taking into consideration the predicted sentiment feature. The aforementioned (ST-SS-REC) process is illustrated in Fig. 1.

The sentiment recognized feature is converted into text and concatenated at the beginning of the tweets. They are separated in BERT by the "[SEP]" token.

Figure 4. BERT sensitivity classification in DTT dataset using tweet only (ST-REC) and tweet with the sentiment (ST-SS-REC).

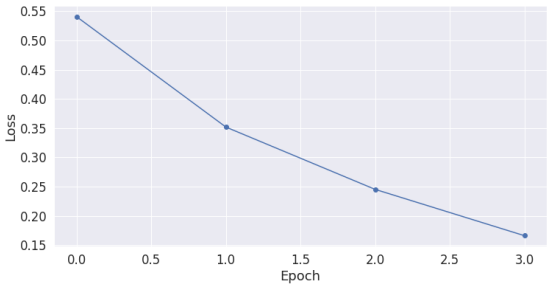


Figure 5. Training loss over one run in SS-ST-REC.

In the case of ST-REC, we decided on a maximal tweet size of 128 characters, whereas, in ST-SS-REC, we used 160 characters. Depending on the original size of tweets, they were either truncated or padded to fit the aforementioned fixed sizes. The decision behind adding 32 more characters in ST-SS-REC is to make up for the introduced sentiment sentence. In ST-SS-REC, each tweet starts with "This is a x tweet" sentence (where, $x \in \{sad, joyful, angry, fearful, surprised, disgusted\}$) and follows with its original text. This way, we make sure to have about the same amount of original tweet characters in ST-REC besides the introduced sentiment sentence.

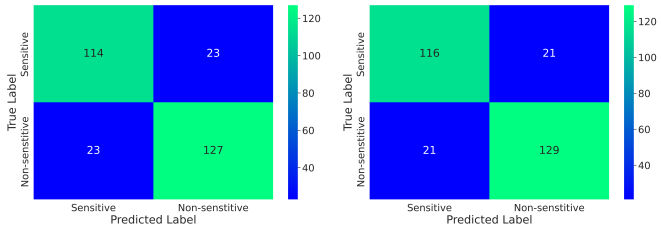
As introduced in Sec. 5, we use four epochs in our evaluations. The average loss over training data is reduced significantly at each epoch, reaching a plateau at the fourth epoch,

as shown in Fig. 5 and the accuracy in validation sets gets higher accordingly.

Our ST-REC model's performance over five runs (shown in Fig. 4) reached the f-1 score of 83,96% and MCC of 67,72%, whereas in the case of ST-SS-REC model, it reached the f-1 score of 87,01% and MCC of 69,83%. Thus, we conclude that the inclusion of sentiment recognition in sensitivity recognition indeed enhances the prediction score.

In comparison, we also observe that our ST-SS-REC model performs better than GloVe+LSTM model used from Wang et al. [27] over the same DTT dataset. Wang et al. [27] presented a GloVe+LSTM model that reached the f-1 score of 85.00% over DTT dataset, i.e, 2,1% lower f1-score than our approach. This however is expected as transformer based models perform better in downstream tasks in contrast to LSTM based classification models.

In Fig. 6, we show the confusion matrices of standard sensitivity model (Fig. 6a) and the integrated sentiment features along with the tweets model (Fig. 6b). In this random evaluated run, we can observe the shift of four tweets (two of them *sensitive* and two others as *non-sensitive*) from being inaccurately predicted in ST-REC to being accurately predicted by ST-SS-REC model.



(a) ST-REC confusion matrix **(b)** ST-SS-REC confusion matrix

Figure 6. Confusion matrix for sentiment recognition over one random run.

6 Limitations and Future Work

We further outline some of the limitations and alternatives in terms of fine-tuning, model sizes, and possible integration of the models into a privacy assistant.

SS-REC model could be further improved by exploring other variations of sentiment classes and performing further hyper-parameter tuning. In turn, this could be reflected to an even more accurate prediction of sentiment in sensitivity datasets, and ergo supposedly higher sensitivity accuracy using ST-SS-REC approach. Furthermore, a balanced dataset with even more diverse emotions would be an interesting direction to explore. For such a dataset, particularly challenging would be to filter tweets by one emotion only, while also keeping a balance between the labelled emotions.

Moreover, the DTT dataset that we used for sensitivity recognition task is rather small. The performance of ST-REC and ST-SS-REC could be further improved with a larger set of qualitative labels.

Another aspect to be considered is the introduction of more recent transformers (e.g., GPT-3 [4] from OpenAI and open-source counterparts, i.e., GPT-Neo [3], GPT-J [24], GPT-NeoX [2] from EleutherAI). They could most probably lead to even better results. However, we decided to use BERT based classification models due to their smaller model size. Besides improving the fine-tuning performance, our ultimate goal is to deploy the models on-device in a privacy assistant mobile application that would leverage the predictions from the models and provide sharing suggestions upon a tweet being shared. Therefore, for such an assistant, smaller model sizes are crucial. The aforementioned transformers have considerable higher sizes (up to 40 GB in the case of GPT-3 and GPT-NeoX). Deploying them on-device is highly unfeasible if not impossible. The assistant would have to rely on cloud based services if we were to apply such huge models. Such approach, however, would introduce other privacy and trust implications. We are particularly working on integrating the models into a mobile user interface that provides sharing suggestions based on the content to be shared.

7 Conclusion

Users keep sharing more and more content online. This leads to possible privacy violations depending on their emotional state. In this work, we investigate the sensitivity recognition and in particular the role of sentiment recognition, as an additional learning parameter, in sensitivity recognition. Thus, we outline two approaches, namely, (1) recognition of sensitive and non-sensitive information using text alone, and (2) recognition of sensitivity given the sentiment inference derived from sentiment recognition model, along with the corresponding text. By evaluating both of the approaches, we demonstrate that the latter one performs better with about a 3% difference (i.e., from 83.96% to 87.01% f-1 score). We further draw a link between six basic emotions and their tendency on privacy exposure.

Acknowledgments

This project is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) and is referenced with the number #317687129.

References

[1] José Alemany, Elena del Val, and Ana García-Fornes. 2020. Empowering Users Regarding the Sensitivity of their Data in Social Networks through Nudge Mechanisms. In *53rd Hawaii International Conference on System Sciences, HICSS.*

[2] Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An Open-Source Autoregressive Language Model. (2022).

[3] Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. *GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow.* https://doi.org/10.5281/zenodo.5297715

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]* (2020). http://arxiv.org/abs/2005.14165

[5] 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.*

[6] Alexandra Chronopoulou, Aikaterini Margatina, Christos Baziotis, and Alexandros Potamianos. 2018. NTUA-SLP at IEST 2018: Ensemble of Neural Transfer Methods for Implicit Emotion Classification. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis.*

[7] Statista Research Department. 2020. Number of Twitter Users Worldwide From 2019 to 2024. Link: https://www.statista.com/statistics/607716/worldwide-artificial-intelligence-market-revenues/. Accessed in: 01.04.2022.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]

[9] Tamara Dinev and Paul Hart. 2005. Internet Privacy Concerns and Social Awareness as Determinants of Intention to Transact. *International Journal of Electronic Commerce* 10, 2 (2005).

[10] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2012. Improving Neural Networks by Preventing Co-adaptation of Feature Detectors. *CoRR* abs/1207.0580 (2012). arXiv:1207.0580 http://arxiv.org/abs/1207.0580

[11] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-term Memory. *Neural computation* 9 (1997).

[12] Giovanni Iachello and Jason Hong. 2007. End-User Privacy in Human-Computer Interaction. *Foundations and Trends in Human-Computer Interaction* 1 (2007).

[13] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent Neural Network for Text Classification with Multi-Task Learning. arXiv:1605.05101 [cs.CL]

[14] Huina Mao, Xin Shuai, and Apu Kapadia. 2011. Loose Tweets: An Analysis of Privacy Leaks on Twitter. In *Proceedings of the 10th Annual ACM Workshop on Privacy in the Electronic Society.*

[15] Mohsen Minaei, S Chandra Mouli, Mainack Mondal, Bruno Ribeiro, and Aniket Kate. 2021. Deceptive Deletions for Protecting Withdrawn Posts on Social Media Platforms. In *Network and Distributed Systems Security (NDSS).*

[16] Manasi Mittal, Muhammad Rizwan Asghar, and Arvind Tripathi. 2020. Do My Emotions Influence What I Share? Analysing the Effects of Emotions on Privacy Leakage in Twitter. In *IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom).*

[17] Saif Mohammad and Felipe Bravo-Marquez. 2017. WASSA-2017 Shared Task on Emotion Intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis.*

[18] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (2019).

[19] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *ArXiv* abs/1910.01108 (2019).

[20] Manya Sleeper, Justin Cranshaw, Patrick Gage Kelley, Blase Ur, Alessandro Acquisti, Lorrie Faith Cranor, and Norman Sadeh. 2013. "I Read My Twitter the Next Morning and Was Astonished": A Conversational Perspective on Twitter Regrets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.*

[21] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*

[22] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems*, Vol. 27. https://proceedings.neurips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv:1706.03762 [cs]* (2017). http://arxiv.org/abs/1706.03762

[24] Ben Wang. 2021. Mesh-Transformer-JAX: Model-Parallel Implementation of Transformer Language Model with JAX. https://github.com/kingoflolz/mesh-transformer-jax. Accessed in: 01.04.2022.

[25] Qiaozhi Wang, Jaisneet Bhandal, Shu Huang, and Bo Luo. 2017. Classification of Private Tweets Using Tweet Content. *IEEE 11th International Conference on Semantic Computing (ICSC)* (2017).

[26] Qiaozhi Wang, Jaisneet Bhandal, Shu Huang, and Bo Luo. 2017. Content-Based Classification of Sensitive Tweets. *International Journal of Semantic Computing* (2017).

[27] Qiaozhi Wang, Hao Xue, Fengjun Li, Dongwon Lee, and Bo Luo. 2019. #DontTweetThis: Scoring Private Information in Social Networks. *Proceedings on Privacy Enhancing Technologies* 2019, 4 (2019). https://content.sciendo.com/view/journals/popets/2019/4/article-p72.xml

[28] Yang Wang, Gregory Norcie, Saranga Komanduri, Alessandro Acquisti, Pedro Giovanni Leon, and Lorrie Faith Cranor. 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.*

[29] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv:1910.03771 [cs.CL]

[30] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv:1906.08237 [cs.CL]

[31] Lu Zhou, Wenbo Wang, and Keke Chen. 2016. Tweet Properly: Analyzing Deleted Tweets to Understand and Identify Regrettable Ones. In *Proceedings of the 25th International Conference on World Wide Web.*