Exploring the Role of Emotions on Privacy Exposure in Twitter

Marvin Kühn
Institute of Computer Science
University of Göttingen
Göttingen, Germany
marvin.kuehn96@stud.uni-goettingen.de

***Abstract*—The goal of this work is to analyze the correlations between emotional sentiments and privacy exposure on tweets. The analysis is based on results of a machine learning text-based Classifier, which we constructed with the PyTorch framework and a Hugging Face Transformers module called BERT. For this task the WASSA-dataset was used to train the model. After cleaning the data and preparing it for the learning task, we used the trained model to predict emotional sentiments for tweets in a separate dataset. In that dataset we incorporated the predicted emotions into the tweets itself, with the purpose to re-train the model with the newly acquired sentiments to predict whether the tweet will be private or public and comparing this extended model to a standard model without the sentiments. The extended model achieved an accuracy of $\sim$90%, over the $\sim$86% from the standard model. An assumption can be made, that the extended model performs better and has potential to be even more precise with some fine tuning.**

***Index Terms*—privacy, twitter, nlp, emotion classification**

I. Introduction

In the world of social media, Twitter is one of the biggest names to exist to date. Contrary to other social media websites like Instagram or TikTok, Twitter defines itself by not necessarily needing good visual content to reach a lot of people, but by expressing themselves through text. Twitter has been known to attract eccentric and emotionally intense opinions [1], but how expressive of your emotions can you be on a public website without compromising your privacy? The research of Wang et al. have examined the privacy sensitivity of tweets [2]. In that paper it was established, that privacy is not only related to personal information, but accumulates through different parameters (e.g. tweeting about location/acquaintances) which are unified in the PrivScore. The PrivScore is generated by a survey, where users give their opinions on the privacy exposure on tweets. Based on this PrivScore, many users compromised their privacy without them knowing. Apart from the obvious privacy exposures like tagging friends, one interesting observation has been that emotions can affect the users choice of words, potentially in an exploitable way.

As sentiment analysis has been a thriving research field in text-based machine learning applications and due to the text-heavy nature of Twitter, a lot of studies and data have been accumulated over time [3][4]. Especially the study of Mittal et al. has conducted similar research on this topic and combined sentiment analysis with privacy concerns [5]. By extending the emotional parameter range of Mittal et al., the following questions are researched in this work:

- To which degree are Twitter-Users emotional sentiments correlated to their privacy exposure?
- And in return: Can we infer concerns for privacy exposure by trying to predict sensitivity levels through the emotional sentiment of a tweet?

This work will begin in *Chapter II* by explaining the fundamentals of a classification process and defining our metrics of measurement. Additionally relevant related work will be presented. In *Chapter III* the conception of the model will be explained, as well as the data preparation needed to start the training. Furthermore evaluation of the models will be shown briefly. In *Chapter IV* the results will be interpreted and presented. *Chapter V* will show limitations of this work and ideas for improving the model. At last there will be a conclusion of this work in *Chapter VI*

II. Foundation and Related Work

To understand the implication of privacy exposure, there needs to be an explanation on when a text can compromise ones privacy. The basic premise of the tweets from the datasets is, that they are all posted publicly. In theory it is possible to completely disable your account from posting publicly, meaning only selected people (friends or followers) are able to see the tweet [6]. However, private accounts are not susceptible to privacy exposure from outside of their own controlled circle and thus not studied in this work. To further understand how private information may be leaked unwillingly by users, a great example is the work of De Capitani di Vimercati et al., which shows that apart from unique ”Identifiers”, there can be ”Quasi-Identifiers” (e.g. sex, location) and ”Confidential attributes” (e.g. a specific disease) [7]. Because our examined dataset is based on the works of Wang et al., we will have to use their evaluated PrivScore, which has been generated by a survey and uses similar metrics to the ones described before [2]. In that survey people had to allocate values based on their perceived sense of privacy exposure from the given sample tweets. This PrivScore will determine if a tweet discloses information to a degree in which privacy has been compromised and is labeled as such in the given dataset.

To complete the foundation of this work, we will now look into the research field of sentiment analysis as we want to determine which sentiments will be considered for this work. Sentiment analysis in NLP-programs has been a research topic in a lot of papers, especially for twitter. Traditional classifiers like Naive Bayes, Maximum Entropy, SVM or word2vec have been used widely [8][3]. In the works of Kanakaraj & Guddeti an ensemble classifier has been used, which outperformed traditional classification methods by 3-5% [9]. In that work traditional methods have been criticized of having certain biases towards specific labels. Mittal et al. conducted research on our to be examined dataset and found that the usage of BERT has been one of the most accurate ones [10][5]. Devlin et al. introduced and also compared the effectiveness of different classifiers like ELMo, GloVe and GPT and found that BERT achieved state-of-the-art results [11]. BERT is short for "Bidirectional Encoder Representations from Transformers" and is a language representation model published by Google [12]. BERTs additional innovation has been the approach of a bidirectional training of Transformers. To elaborate on the bidirectional training: Traditional language models either read the data from left-to-right or vice-versa. BERT is reading the data all at once, making it bidirectional or even non-directional [13]. It can be used for a variety of things like name recognition, sentence completion and question answering. However for our task only (sentiment/sensitivity) classification will be relevant. When we use the BERT model, it will be pre-trained. This means that the model has undergone unsupervised training based on the BooksCorpus (800M words) and english Wikipedia (2500M words) which will give us an advantage over untrained models [11]. To make use of the pre-trained model we need to "fine-tune" it. Fine-tuning the BERT model will allow us to supply it with prepared data, turning it into a supervised machine learning task. This will inherently mean that data inspection and data preprocessing will be necessary to fine-tune the model effectively [14] . Furthermore, fine-tuning includes the optimization of hyper-parameters like learning rates and batch sizes [11].

At last we need to define the classification of emotions. There are many models which use the 4 primary emotions (sad, anger, fear, joy) [15][5], which have also been used by Mittal et al. However, as we want to expand the work of Mittal et al., two emotions will be added. The research of Ménard et al. and Ekman & Friesen has shown that the addition of "surprise" and "disgust" can be beneficial to the information conveyed [17][16]. The given dataset already has these six emotions labeled. This means that we fulfill the prerequisite to execute a NLP supervised machine learning strategy.

III. Model Design

This chapter illustrates the choice, conception and data preparation needed to train and evaluate a machine learning algorithm for the task of classifying tweets. We will first have to define the scope of our classifier to pick a suitable framework and model. We want to measure the accuracy and the confusion maps of the classifiers to interpret the results. It would be favorable to achieve a high accuracy (minimal: 60%, optimal: 70% and upwards) in our first model to make sure that our sentiment predictions are correct most of the time for the second dataset. This is a basic requirement, which must be fulfilled to conduct this research.

A. Architectual Choice

Based on the accuracy results of Mittal et al., the transformers bundle from Hugging Face provides a powerful interface to use BERT, which has achieved state-of-the-art results in NLP applications [11]. The BERT language model has many optimized sub-models for specific tasks. As all our models will be classifiers, BertForSequenceClassification will be used as a regression based machine learning model. The underlying needed framework is PyTorch. Another option which would be interesting for future work would be the GPT-Neo model, which has performed close to the top-level GPT3 model [18]. As there is no physical PC available with the needed computing power to run this machine learning task in an efficient time window, third-party applications like Amazon Sagemaker Studio and Google Colab are used in this work. As these tools are restricted in their time usage, BERT is a good fit due to its accessible and easy to use interface.

B. Data Inspection

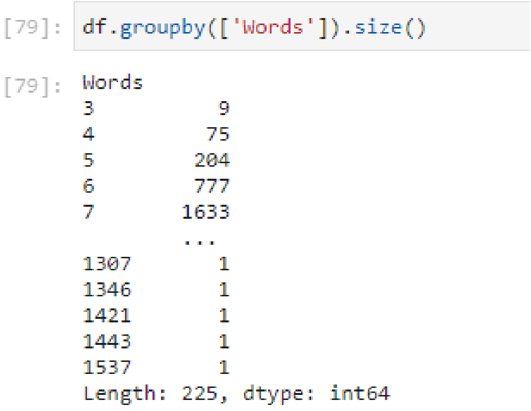
To train our data efficiently, we will need to inspect our datasets. For the WASSA-dataset there are ∼150,000 tweets with their six corresponding emotions attached to them. The second dataset from the DontTweetThis-Paper (DTT) [2] contains ∼3000 tweets with their corresponding privacy label.

WASSA will be a multiclass-classification, containing six emotions (anger, sad, fear, joy, surprise, disgust), which means there will be two additional labels to extend the work of Mittal et al. based on our findings in *Chapter II* [5]. Various models with satisfying results have been trained on this dataset already [19][20]. The amount of tweets for each emotion is generally balanced, as we have a mean of around 25064 tweets per emotion with a maximum deviation of 7.5% to the lower end (sad-emotion) and 11.5% to the top end (joy-emotion), while the other emotions are within 1.5% of the mean. This data split is sufficient for the training if we stratify the values accordingly in the training and validation set.

DTT will be a binary-classification containing the information on whether the tweet is considered a private or public tweet. The dataset is split in half in regards to the sensitivity sentiment. While this is an optimal binary split, the size of data (2900 entries) might be a limiting factor in this research, as it plays a crucial role in increasing the accuracy for text-based classifiers in a natural way [21]. The significance of predictions from the BERT model has to be critically examined.

C. Preprocessing and Preparing Data

The next step is to look at tangible data to determine if data cleaning is necessary. Later we will need to tell the Tokenizer



```
[79]: df.groupby(['Words']).size()

[79]: Words
      3          9
      4         75
      5        204
      6        777
      7       1633
               ...
      1307       1
      1346       1
      1421       1
      1443       1
      1537       1
      Length: 225, dtype: int64
```

Fig. 1. Occurrence of amount of words in a tweet

how many words are used per tweet at the maximum. Because tweets have a hard limit of 280 characters as of 2022 [22], there must be a cap on the amount of words in a tweet (at least if somewhat real sentences are used). Even though the model is equipped to truncate the length of the tweets, the optimal training will be achieved if every tweet is used to its full length. As we check the WASSA-dataset for the length of the words as seen in Fig. 1, there seem to be abnormally long tweets. By investigating the specific indices of the occurring anomalies, we can isolate two causes for the long tweets:

1) The character (”) used only once causes the iterator to not detect the end of a dataset row, because it assumes it is inside a quote from a string and will continue further, until another single use of (”) finally closes the string.
2) The tagging of users with (@USERNAME) doesn’t seem to count towards the set cap of 280 characters. But will render the tokenized data useless, because it truncates from the end to the start.

The first instance is not to be ignored, not only because it makes the starting tweet unfavorable for the classifier, but because it restricts a potentially enormous amount of data to be processed by truncating it altogether. To solve this problem all occurences of (”) characters were removed. For the second instance we had to take into consideration, that the amount of (@USERNAME) used could correlate to a specific emotional sentiment. By further investigation it was assessed that the amount of times these (@USERNAME) filled tweets occurred was manageable. Furthermore the mass-tagging is located at the start of the tweet almost always. Because truncation happens at the end of a tweet, the decision was to cut the mass-tagging at the start manually to an amount where the longest tweets of mass-tagging are about the same length as legit tweets, which in practise means a maximum of ∼5 taggings per tweet at the start was allowed.

The last step for the WASSA-dataset was to replace occurences of (”http://url.removed”) with just (”URL”), because it has no additional information over the replaced value and will shorten the average length of one entry. The main issue for most datasets is that there is no norm to how the data is presented. As one dataset may visualize URLs in a different way than another dataset, the best practise would be to normalize the data over the given datasets. However this option would not work if a general model will be built to use for uninspected datasets. The options which arise would be to do nothing or try to match as many representations as possible. While further data cleaning could potentially help the classifier, Jianqiang & Xiaolin concluded, that pre-processing URLs, stop words and numbers is appropriate to reduce noise in the dataset, but only minimally affects the performance of classifiers [23]. Another possible procedure is to process emoticons, as they are used consistently throughout the dataset. Emoticon processing has been used for the WASSA-dataset, as seen in the works of Chronopoulou et al. [24][25], but is not utilized for this work.

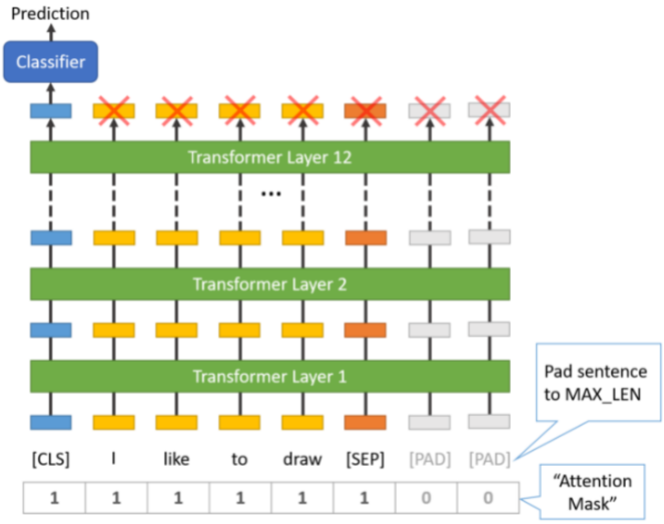


Fig. 2. BertTokenizer visualization [27]

The DTT-dataset has no such irregularities based on the same investigation. No further data preparation was needed for this dataset.

To further prepare the data for the modeling we will need to determine the size of training and validation splits of the tweets. A validation size of 15% was chosen for both datasets. Furthermore, because we are facing a multiclass-classification for the sentiment training, the six emotions needed to be transformed into numerical values from 0 to 5. Again, for DTT the data is available in the right shape and no further transformation is needed.

D. Modeling

In the modeling phase, most of the optimization (this includes hyperparameter optimization) takes place. The first step is to tokenize every tweet with the BertTokenizer (see Fig. 2). This will split up the words and symbols of the tweets into an array of individual IDs in a way that the same words will have the same ID over the dataset. Additionally there are special tokens, which specify the start, the end and the padding

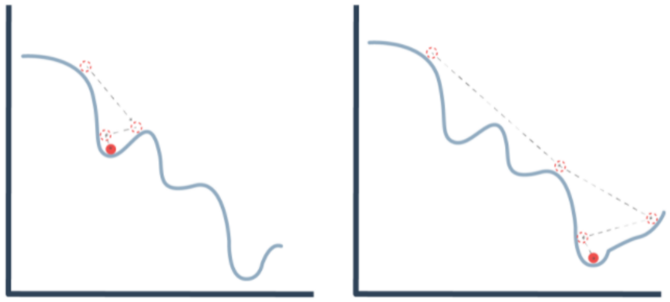


Fig. 3. Learning rate effect [26]

of every tweet. The padding will be ignored by the classifier through the attention mask.

The batch size indicates how many samples (here: tweets) are used before the internal model parameters are updated. Research and testing has shown that a batch size of 16 or 32 are optimal [27]. For our model we will chose 16. The next parameters will be the learning rate (lr), the epsilon value (eps) and the amount of epochs trained. Generally the epsilon values task is to stop divide by zero errors resulting the model to get stuck, however choosing a big value will cause the weight updates to be less significant, thus risking a slower learning progress. The learning rate directly dictates the magnitude of change for the weights. Choosing a very low learning rate may result in the classifier getting stuck at a local minimum instead of the overall minimum (see Fig. 3). The amount of epochs will tell the classifier how many times it should iterate over the whole dataset. Because we are saving the model after each epoch, we can choose a bigger value at first and examine when the epochs might overfit. In our case the model started to overfit at around epoch 4 most of the time. After some additional research [28] and internal testing, the best models were thus achieved with the following parameters:

batch size: 32 **eps:** 1e-8 **lr:** 2e-5 **epochs:** 3

An important note is that the same training splits and the same hyperparameters are set for both datasets. This might not be optimal for the standalone accuracy of the DTT-dataset, but it will ensure a fair premise to compare the results later.

E. Training and Evaluation

For our sentiment prediction we achieved an overall accuracy of about 71%, with the confusion matrix shown in Fig. 4. The accuracy values for specific classes are as follows:

Sentiment-Label	Accuracy
anger	65%
disgust	66%
fear	78%
sad	70%
surprise	68%
joy	80%
Overall	71%

In *Chapter III: B* we assessed, that joy-tweets have the most entries in our dataset, which correlates with the predictions of the classifier. However, the lowest amount of entries were sad-tweets, which do not correlate with the predictions of the classifier. Thus we can rule out that our classifier learned a strategy solely based on label count.

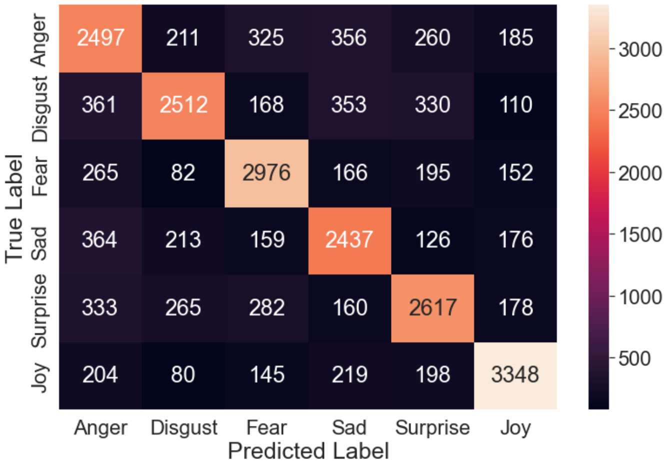


Fig. 4. Confusion Matrix of the sentiment model

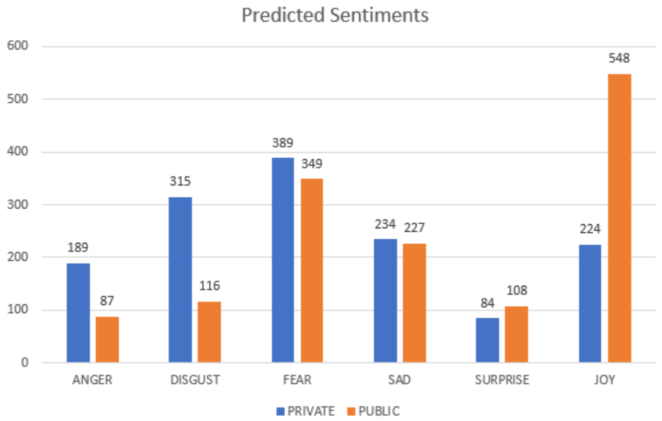


Fig. 5. Sentiment prediction on the DTT-dataset

Continuing our strategy, our sentiment classifier was used to label emotions onto the DTT-dataset. The result is shown in Fig. 5. Inherently it is evident that joy is an emotion which is shared publicly, while negative sentiments, especially disgust and anger are correlated more with private tweets. The next step will be to incorporate the text into the dataset. The following table shows an example of how tensors are prepended with the sentiment and then reshaped throughout the dataset:

Old Tensor		
Text	Sensitivity	Emotion
tummy ache is over:D smoo..	0	5 (Joy)

New Tensor	
Text	Sensitivity
This is a joyful tweet. tummy ache is..	0

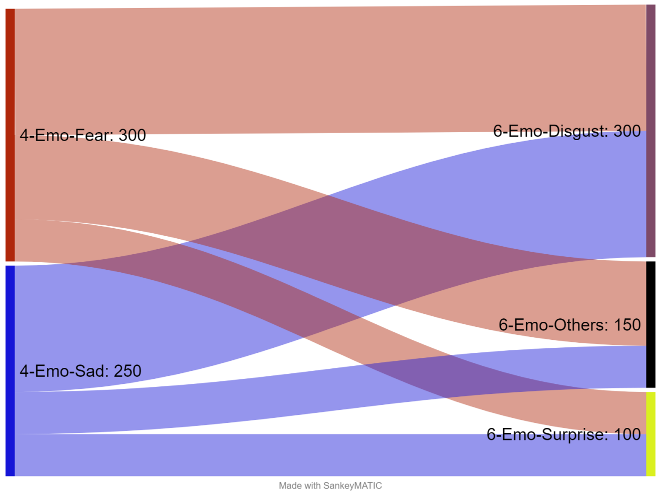
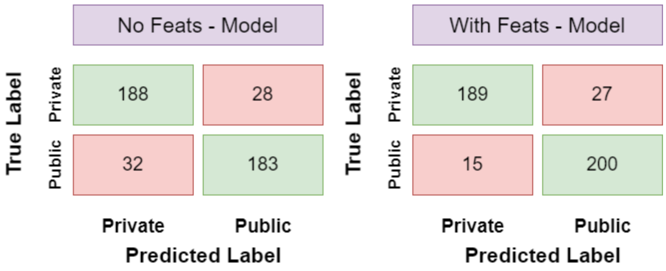


Fig. 6. Possible migration of private tweets from "fear" and "sad" to the new labels



True Label	No Feats - Model: Predicted Private	No Feats - Model: Predicted Public	With Feats - Model: Predicted Private	With Feats - Model: Predicted Public
Private	188	28	189	27
Public	32	183	15	200

Fig. 7. Confusion matrices for sensitivity prediction models

The last step will be to train and evaluate two identical models, one on our newly generated dataset with sentiment feats incorporated and one on the basic dataset without any additional information. For the basic model, an overall accuracy of 86% has been achieved, with around 87% accuracy for private and 85% for public tweets. The extended model achieved an overall accuracy of 90%, with around 87.5% accuracy for private and 93% for public tweets. The implications of the these measurements will be analyzed in the following chapter.

IV. Results

Analyzing the results chronologically, the confusion matrix of the sentiment classifier (see Fig. 4) shows the strongest predictions for the emotions "joy" and "fear". If we try to classify the six emotions into positive and negative emotions, only "joy" fits the positive category fully, with "surprise" being neutral. All of the other emotions would be classified as negative. As simpler sentiment classifiers only predict between positive and negative tweets [29], "joy" being the only true positive emotion can correlate with the classifier having a higher accuracy for that label.

If we compare the distribution of emotions in respect to their sensitivity label in Fig. 5 with the same diagram from Mittal et al., "joy" and "anger" appear to stay in the same relation strongly, while "sad" has changed the most with fear being close behind [5]. Looking at the two new emotions, "disgust" is correlated the most to privacy exposure in relative values. "Surprise" is balanced between privacy and public. From this we can deduct, that the emotion "disgust" has drawn mostly from the private segments of the emotions "fear" and "sad", while "surprise" has no explicit emotion from which it has drawn particularly, a visualization of the migration can be seen in Fig. 6. The general distribution of emotions in regards to privacy exposure could be interesting to look at in the sense of information gain. If we assume a classifier might not be able to differentiate between sad public tweets and sad private tweets, then it has nothing to gain from the information that it is a sad tweet. In contrary if we apply the same assumption to a joyful tweet, the information gain from the sentiment label could nudge the classifier in the right direction. This implies the possibility that it might be a better practice to evaluate which specific emotions have a higher probability of changing the sensitivity label of a tweet, as the disgust emotion has shown. The main difficulty for this task is to prevent the classifier to rely only on the sentiment. This could be tackled with the help of weights or adjusting the label based on the confidence level of the sentiment classifier. Furthermore "anger" and "disgust" can be identified as the sentiments which compromise the privacy the most. As deducted before, a positive emotion does not seem to correlate with a compromise of privacy, so why don't negative emotions split up their sensitivity score equally? An assumption can be made that "fear" and "sad" are dealt with in an introverted or "flight" type of way. In contrary the emotions "anger" and "disgust" are dealt with in a more extroverted or "fight" type of way, meaning that the user will act more irrationally and have a higher probability of sharing something that could be classified as private information.

In the second section of this work, the evaluation has shown that the overall accuracy is apparently better with the extended model, however one of our main goals is to check if privacy exposure takes place. This means that a correctly predicted public tweet is not worth as much as a correctly predicted private tweet. Comparing the more detailed accuracy shows, that the models perform basically identical (87% vs 87.5%) in regards to predicting privacy exposure. This is also visualized in in Fig. 7. That said, there seems to be no trade-off happening, which means the model with feats appears to perform the same as the basic model at the minimum and marginally better at the optimum. The transformation of the confusion matrix is also worth noting. Because the accuracy of private predictions stayed the same, the rising accuracy in correct public predictions is drawn exclusively from the incorrect labeling of public tweets. The reduction of the type 1 error is not as desired as reducing the type 2 error in our case.

To come back to the overall accuracy, the extended model still showed a meaningful rise. Even if the sentiments only helped predicting more public tweets correct, it is still an improvement. This inherently means that it can be confidently said, that the knowledge of emotional sentiments do influence the accuracy of the classifier in a positive way. For other iterations it could be interesting to see if it would be possible to optimize the confusion matrix towards the precision of correctly predicted private tweets, as this could have viable use cases for warning users from tweeting sensitive information like Wang et al. has proposed [2] and reducing the type 2 error in our case. As the internal hyperparameter optimization was done by trial and error, there were some iterations where the prediction accuracy of private tweets has been significantly higher than the public predictions. Sadly in these occasions a lower overall accuracy was achieved, so that the absolute amount of correct prediction has been to low to be shared in this work. However this could imply that it might not be unfeasible for the classifier to predict privacy exposure better with a minimal change in the data or model.

V. Limitations and Future Research

This research was limited mainly by the needed processing power to train the models. Computing on a private PC is not feasible for the size of this dataset. By using third-party applications like Amazon Sagemaker and Google Colab, the training could be executed fully. However, these applications limit the time of available GPU processing power for users to a few hours per session. This means that optimization and "playtesting" different hyperparameters could not be realized to an extent which would be more satisfying.

To elaborate on the training, the model is currently overfitting fast to an extend which should be investigated in further research. Measures against overfitting like regularization, adding more data or simply testing different hyperparameters could potentially improve the accuracy of the classifier and increase the quality of the comparison done in *Chapter IV*. Especially for the DTT-dataset, the amount of available tweets was low to an extend that the model overfitting could not be handled accordingly.

Another interesting approach would be to check if some sort of majority label prediction takes place in our second classifier. As we have seen in Fig. 5, the emotion "joy" is found overwhelmingly in public tweets. This could explain why our classifier was more accurate for public tweets.

As the exploration of privacy exposure through emotions has been promising, even more things like emoji/punctuation-sign usage, tweet length, picture/video usage or even overarching parameters like time of day or locations (if available) could all affect users to generally behave differently and allowing classifiers to see patterns more easily. For future research these parameters could be incorporated into the dataset as well to examine if they play a role in privacy exposure.

VI. Conclusions

This work has expanded the works of Mittal et al. [5], to incorporate the emotional sentiment of tweets into the dataset itself and check if the additional information helps in predicting privacy exposure. After initially preparing the data and predicting the emotional sentiments, we achieved an accuracy of $\sim$71%. This sentiment model was used to extend the data of the second dataset with emotions to again train two models based on that data, achieving $\sim$86% for the basic and $\sim$90% for the extended one.

Our results have shown that the accuracy rose slightly, which should be an indicator that the added information is not negatively affecting the classifier. However the small range in which the accuracy has changed is not at a level at which certain increase could be declared universal. Furthermore it was assessed that the improvement of the accuracy has not reduced the type 2 error with the same strength. Based on the premise of this work to detect privacy exposure, this means that further research is needed to improve the model. There were many identified possibilities to improve the classifier of privacy exposure. These reach from improving the preprocessing of data to further examining the role of specific emotions like "anger" and "disgust", which seem to correlate with privacy exposure more strongly.

The next goal would be to increase the general accuracy of the classifier and apply other parameters into the dataset, providing more options to quantify if these parameters affect the privacy exposure of twitter users.

References

[1] Constance Duncombe, The Politics of Twitter: Emotions and the Power of Social Media, International Political Sociology, Volume 13, Issue 4, December 2019, Page 417

[2] Wang, Qiaozhi & Xue, Hao & Li, Fengjun & Lee, Dongwon & Luo, Bo. (2019). #DontTweetThis: Scoring Private Information in Social Networks. Proceedings on Privacy Enhancing Technologies. 2019. 72-92.

[3] Imran, Muhammad & Mitra, Prasenjit & Castillo, Carlos. (2016). Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages.

[4] S. Piao and J. Whittle, "A Feasibility Study on Extracting Twitter Users' Interests Using NLP Tools for Serendipitous Connections," 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, 2011, pp. 910-915

[5] Mittal, Manasi and Asghar, Muhammad Rizwan and Tripathi, Arvind, Do My Emotions Influence What I Share? Analysing the Effects of Emotions on Privacy Leakage in Twitter (October 30, 2020). The University of Auckland Business School Research Paper

[6] Who can see your tweets – twitter privacy and protection settings. Available at: https://help.twitter.com/en/safety-and-security/public-and-protected-tweets (accessed 09.03.2022)

[7] Vimercati, Sabrina & Foresti, Sara & Livraga, Giovanni & Samarati, Pierangela. (2012). Data privacy: Definitions and techniques. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 20. pp. 793-817.

[8] S. A. Phand and J. A. Phand, "Twitter sentiment classification using stanford NLP," 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM), 2017, pp. 1-5

[9] M. Kanakaraj and R. M. R. Guddeti, "Performance analysis of Ensemble methods on Twitter sentiment analysis using NLP techniques," Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015), 2015, pp. 169-170.

[10] WASSA-2017 Shared Task on Emotion Intensity (EmoInt) Available at: http://saifmohammad.com/WebPages/EmotionIntensity-SharedTask.html (Accessed 03.03.2022)

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, & Kristina Toutanova (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR, abs/1810.04805.

[12] BERT: Bidirectional Encoder Representations from Transformers, Available at: https://huggingface.co/docs/transformers/model$_$doc/bert/ (Accessed 03.03.2022)

[13] Rani Horev, BERT Explained: State of the art language model for NLP, Available at: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270 (Accessed 09.03.2022)

[14] Xiaobing Sun, Xiangyue Liu, Jiajun Hu, and Junwu Zhu. 2014. Empirical studies on the NLP techniques for source code data preprocessing. In Proceedings of the 2014 3rd International Workshop on Evidential Assessment of Software Technologies (EAST 2014). Association for Computing Machinery, New York, NY, USA, 32–39.

[15] Vora, Parth & Khara, Mansi & Kelkar, Kavita. (2017). Classification of Tweets based on Emotions using Word Embedding and Random Forest Classifiers. International Journal of Computer Applications. 178. pp. 1-7.

[16] Ménard, Mickaël & Richard, Paul & Hamdi, Hamza & Daucé, Bruno & Yamaguchi, Takehiko. (2015). Emotion recognition based on heart rate and skin conductance. PhyCS 2015 - 2nd International Conference on Physiological Computing Systems, Proceedings. 26-32.

[17] Ekman, P. & Friesen, W., 1982. Measuring facial movement with the facial action coding system. In: Emotion in the human face (2nd ed.). s.l.:New York: Cambridge University Press.

[18] Matthew McAteer (April 2021), Messing with GPT-Neo, Available at: https://matthewmcateer.me/blog/messing-with-gpt-neo/ (Accessed 07.03.2022)

[19] Tafreshi, S., De Clercq, O., Barriere, V., Buechel, S., Sedoc, J., & Balahur, A. (2021). WASSA 2021 Shared Task: Predicting Empathy and Emotion in Reaction to News Stories. In Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (pp. 92–104). Association for Computational Linguistics.

[20] Klinger, R., De Clercq, O., Mohammad, S., & Balahur, A. (2018). IEST: WASSA-2018 Implicit Emotions Shared Task. In Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (pp. 31–42). Association for Computational Linguistics.

[21] Jayme Garcia Arnal Barbedo (2018). Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. Computers and Electronics in Agriculture, 153, 46-53.

[22] Counting characters when composing Tweets, Available at: https://developer.twitter.com/en/docs/counting-characters#:~:text=In%20most%20cases%2C%20the%20text,as%20more%20than%20one%20character., (Accessed 03.03.2022)

[23] Z. Jianqiang and G. Xiaolin, ”Comparison Research on Text Preprocessing Methods on Twitter Sentiment Analysis,” in IEEE Access, vol. 5, pp. 2870-2879, 2017.

[24] Chronopoulou, A., Margatina, A., Baziotis, C., & Potamianos, A. (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 (pp. 57–64). Association for Computational Linguistics.

[25] Alexandra Chronopoulou, ntua-slp-wassa-iest2018, GitHub Link Available at: https://github.com/alexandra-chron/ntua-slp-wassa-iest2018/blob/master/utils/nlp.py (Accessed 04.03.2022)

[26] AdamW: When to change optimizer & optimizer parameters, Available at: https://peltarion.com/knowledge-center/documentation/modeling-view/run-a-model/optimizers/adamw (Accessed 08.03.2022)

[27] Chris McCormick, BERT Fine-Tuning Tutorial with PyTorch, Available at:https://mccormickml.com/2019/07/22/BERT-fine-tuning/ (Accessed 05.02.2022)

[28] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, & Veselin Stoyanov (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach.

[29] T. Ghorpade and L. Ragha, ”Featured based sentiment classification for hotel reviews using NLP and Bayesian classification,” 2012 International Conference on Communication, Information & Computing Technology (ICCICT), 2012, pp. 1-5.