SUSPENDED ACCOUNTS: A SOURCE OF TWEETS WITH DISGUST AND ANGER EMOTIONS FOR AUGMENTING HATE SPEECH DATA SAMPLE

WAFA ALORAINY[1,2], PETE BURNAP[1], HAN LIU[1], AMIR JAVED[1], MATTHEW L. WILLIAMS[3]

[1]School of Computer Science And Informatics, Cardiff University
[2]College of Sciences and Humanities, Shaqra University
[3]School of Social Science, Cardiff University
E-MAIL: alorainyws@cardiff.ac.uk, burnapp@cardiff.ac.uk, liuh48@cardiff.ac.uk
javeda7@cardiff.ac.uk, williamsm7@cardiff.ac.uk

Abstract:

In this paper we present a proposal to address the problem of the pricey and unreliable human annotation, which is important for detection of hate speech from the web contents. In particular, we propose to use the text that are produced from the suspended accounts in the aftermath of a hateful event as subtle and reliable source for hate speech prediction. The proposal was motivated after implementing emotion analysis on three sources of data sets: suspended, active and neutral ones, i.e. the first two sources of data sets contain hateful tweets from suspended accounts and active accounts, respectively, whereas the third source of data sets contain neutral tweets only. The emotion analysis indicated that the tweets from suspended accounts show more disgust, negative, fear and sadness emotions than the ones from active accounts, although tweets from both types of accounts might be annotated as hateful ones by human annotators. We train two Random Forest classifiers based on the semantic meaning of tweets respectively from suspended and active accounts, and evaluate the prediction accuracy of the two classifiers on unseen data. The results show that the classifier trained on the tweets from suspended accounts outperformed the one trained on the tweets from active accounts by 16% of overall F-score.

Keywords:

Data annotation, Hate speech data sets, Classification, Suspended accounts

1. Introduction

In recent years, twitter have provided a panel where people can interact with each other. Although it has become a useful source for information spread, unfortunately it has produced antagonistic contents. However, finding and predicting such information is a formidable task due to a large amount of user-generated contents that are spread on the web. One challenge is to separate high quality contents from offensive, hateful, abusive or massively biased ones using text classification techniques. In order to perform experiments on hate speech detection, having access to labelled corpora is essential. Since there is no commonly accepted benchmark corpus for the task, authors usually collect and label their own data.

The reliability of the human annotations is crucial, both to ensure that the algorithm can accurately learn the characteristics of hate speech, and as an upper bound on the expected performance [1]. A study done by [2] clarified that the agreement of the annotators was very low because they revealed that there is considerable ambiguity in existing definitions. A given statement may be considered hate speech or not depending on people's cultural background and personal sensibilities. This clarifies the problem of the annotation when supervised learning is adopted. [3] compared crowd-sourced annotations achieved using AMT with annotations created by expert annotators, and found large differences in terms of agreement rate. In addition to the reliability, human annotation is costly when the research is aimed to examine big data sets. Furthermore, sometimes people tend to annotate a number of samples (e.g. 2000) but the annotation results become imbalanced (e.g. 1800 for the benign ones and 200 for the hateful ones). Also, the status of hate speech is variable, which means that hateful instances may be considered non-hateful later on. In this case, researchers need to augment their annotated data, while this process is charged.

Currently, the data sources that are used in Twitter hate speech studies include: [4], [5], [6], [1], [7]. Collecting data from Twitter are affected by several factors that might appear during the advanced stages. For example, several studies collected the data by performing an initial manual search of com-

mon slurs and terms used pertaining to religious, sexual, gender, and ethnic minorities for crowd-sourcing process. In the results, the collected data might contain tweets that were annotated differently. To solve this, several studies (e.g.[1]) provided a list of criteria founded in critical race theory, and use them to annotate a publicly available corpus. While this increases the proportion of hateful posts on resulting data sets, it focuses the resulting data set on specific topics and certain sub types of hate speech (e.g. hate speech targeting Muslims). Furthermore, human annotation suffers from the problem of disagreement even though the criteria were specified, e.g. the tweet text "go to your mum" would be annotated differently by different annotators with different cultures.

Another problem is the size of the annotated data set. In general, the size of collected corpus varies depending on specific works on hate speech detection, ranging from around 100 labelled comments [8], [9], to several thousand comments used in other works, such as [1]. Another reason for the size differences lies in the simple fact that annotating hate speech is an extremely time consuming and costly task, which might end with the case that there are much fewer hateful than benign comments present in randomly sampled data, such as data sets collected in [6], and therefore a large number of comments have to be annotated to find a considerable number of hateful instances.

In this study, we aim to find a robust source of hateful contents without relying on manual annotation. The idea came up when we noticed that the annotated hateful tweets produced by suspended accounts (which violated twitter policy and were reported by other Twitter users) contained more antagonistic languages than the hateful tweets produced by active accounts.

According to twitter suspension policy, users may not use Twitter for the purposes of spamming yet there are different reasons for account suspension that include: 1- account security at risk e.g., an account has been hacked or compromised; 2- abusive tweets or behavior e.g., sending threats to others or impersonating other accounts [10]. The point here is that there are accounts which are suspended due to producing hateful contents. Twitter provide an option for people to report an account that produces violence, and they were asked to provide several tweets from that account to better understand the context. We assume that the suspended accounts involved in a specific related event (e.g. religion) produce contents that include strongly violent, abusive or hateful languages.

The above points suggest that suspended accounts are reliable source of hateful tweets, since these accounts kept posting tweets in the aftermath of hateful events (related hash tag or keywords) and were thus suspended by Twitter. In other words, when collecting tweets in the aftermath of the Woolwich attack, the tweets from suspended accounts would be recognized as either spams or abusive ones. We examined the assumption that text posted from the suspended accounts could be used as reliable source for augmenting an existing hateful data set or creating a new one. Initially, we proposed to analyze three kinds of human annotated tweets: (1) hateful tweets produced by active accounts, (2) hateful tweets produced by suspended accounts and (3) neutral annotated tweets.

The analysis was implemented by characterizing the suspended, active and neutral accounts' text using emotion analysis. This showed that although both active and suspended users produced hateful contents, the text which is produced by the suspended accounts included different and more antagonistic emotional attitudes than the active ones. Then, we examined the effectiveness of the different characteristics of text from the suspended, active and neutral accounts for predicting hate speech. This was implemented by training a Random Forest (RF) classifier on the semantic features of the tweets from each of the above types of accounts, in order to validate it on an unseen annotated data set which contains neutral and hateful samples. Interestingly, the classifier trained on tweets from suspended accounts performed better than the one trained on tweets from active accounts by 16% in terms of F-measure for both hateful and benign text. The results suggest that the suspended account produced more defined hateful and antagonistic language than the active one. To the best of our knowledge, none of the previous studies considered the status of twitter accounts as a pointer of a strongly hateful source towards performing the data collection task based on the account status.

In this paper we will review the related works in section 2. The data sets are explained in section 3 and the conducted experiments are explained in section 4. Our results are presented and discussed in section 5 and we conclude with recommendation in section 6.

2. Related Work

In this section, we explore the studies that have been done for the purpose of distinguishing the suspended accounts than the active ones from the text perspective. In particular, the authors of the studies [10] [11] showed that there are substantial differences between the suspended and active accounts content presented by the style of writing and the network interactions. They found that deleted and suspended accounts use less hashtags, mentions and pincushions. In addition to the linguistic and network features, they found that the suspended accounts produced less anger and fear but more disgust emotions. The previous study suggests that the suspended accounts include more

distinguishable contents than the active and deleted ones. However, their study was not implemented on data sets collected during a specific hate speech related-event so the emotions analysis reflected general but non-hateful speech attitudes.

Another study by [12] considered the account suspicion after spam filtering for characterizing users who provide the hash-tag #Gamegate containing abusive and bullying text. They performed preliminary measurements on how the Twitter suspension mechanism deals with such abusive behaviors and they showed that the accounts that send this hash-tag tend to be suspended (20%) more than to be deleted (13%). [13] introduced a comparison of the suspended accounts measurements between aggressive and cyberbullying. Similar to previous studies, the authors tended to measure the behaviour of suspicion among hateful accounts. However, both of the two previous studies tended to investigate the behaviours of the suspended accounts (which were suspended because of abusive and bullying behaviours) rather than to investigate the tweet text itself.

3. Data

For training the classifier, we collected new tweets from the suspended and active accounts according to the tweets from an annotated data set by [5]. In particular, 2000 out of the collected 450,000 tweets are sampled for human coding. Coders were provided with each tweet and the question: "is this text offensive or antagonistic in terms of race ethnicity or religion?" They were presented with a ternary set of classes- yes, no, undecided. We utilized the CrowdFlower online service that allows for Human Intelligence Tasks, such as coding text into classes, to be distributed over multiple workers. The results of the annotation exercise produced a "gold standard" data set of 1,901 tweets, with 222 instances of offensive or antagonistic content. The 222 hateful samples were divided depending on the account status as follows: 120 from active accounts, 28 from deleted accounts and 74 from suspended accounts. As the size of the hateful sample was very small, we extended this sample by searching extra tweets from the collected 450,000 tweets. The key words for searching extra hateful tweets were extracted from the 222 annotated tweets (e.g. "nigga"). We ended up with 798 tweets that were divided into three sections (Active 370, deleted 51 suspended 377). We considered the contents from active and suspended accounts and discarded the contents from deleted accounts because there are different reasons for accounts to be deleted in Twitter (e.g. personal reasons). For testing our classifier, we use the data set published by [7]. The set of tweets were collected depending on words and phrases that were identified as hate speech by internet users, compiled by Hate-base.org and manually coded by CrowdFlower (CF) workers.

4. Methods

In this section, we describe our methods that were implemented for investigating the effectiveness of the contents from suspended accounts on detecting hate speech. So, we will not aim to use and compare different state-of-the-art methods for feature extraction. In particular, we will present how the data was prepared and how the emotion analysis was conducted. Also, the setup for semantic learning of embeddings and hate speech classification will be specified.

4.1 Data Preparation

The data sets were prepared through status extracting, spam filtering and text pre-processing. In terms of status extracting, we used a Twecoll [1] python tool for investigating the account statuses depending on the error code results (401 and 404 indicate that the account has been suspended and deleted respectively). Also, spam detection and filtration is a complicated process as no certain method would detect spam in 100%. In particular, we considered a tweet as spam if it contained URLs or more than 14 hash-tags according to [14]. In addition, the data sets were cleaned from extra symbols and stop words.

4.2 Emotion Analysis

In this section we aim to identify the eight primary emotions (anger, anticipation, disgust, fear, joy, sadness, surprise and trust) as defined by [15] and also look at the sentiment (negative or positive) in the tweet. To identify the eight primary emotions, we have created an emotion and sentiment detection program that identifies emotion based on keywords and emoticons. The program uses multiple dictionaries containing words and emotions associated with each word. These dictionaries were built using Wordnet Affect Lexicon [16] and dictionaries (NRC-Emolex and Hastag Emotion Corpus) from National Research Council of Canada [17][18]. In addition to these dictionaries, we have used emoticons in a tweet to identify emotion being reflected in a tweet.

While looking for various emotions in a tweet, the program first tries to identify if there were any emoticons in the tweet. If found, it then checks emotions associated with that tweet. In the next iteration, the program cleans the tweet from all stop words, and start checking emotions associated with words and

[1] https://github.com/jdevoo/twecoll

hashtags found in the tweet. The program uses all available resources in terms of dictionaries to identify emotions in the tweet. Finally, we check for sentiment associated with a tweet. Once the iteration is complete, a log file is created reflecting emotion and sentiment identified in the tweet.

4.3 Semantic Learning

We applied the Doc2vec embedding learning approach for transforming words or textual instances directly into feature vectors, through training of deep neural networks. In general, text embedding is aimed at learning numerical representation of words, sentences or even more complex textual instances, which can lead to reduction of the dimensionality and sparsity of feature vectors, in comparison with other feature extraction methods, such as Bag-of-Words (BOW) and N-grams (NG).

Since each textual instance (document) consists of a list of words, a document vector can be trained alongside the corresponding word vectors. We applied Doc2vec according to the study done by [9], who validated that the Doc2vec representations are meaningful and that semantically similar words are close to each other in the embedding space (since we intended to examine different sets of tweets (sentence) from suspended and active accounts). Document embedding can be achieved through two core models: distributed memory (DM) and distributed bag-of-words (DBOW). While the DBOW model has performed well for detecting the hateful samples (e.g. it detects all the hateful samples but only 103/3269 of benign samples), it failed in detecting the benign samples (detect low number). We decided to apply Distributed memory (DM) model for embedding learning because it produces better features which performed well for both hateful and benign prediction. We learned vector representations with 600 dimensions and the context of windows size of 2.

4.4 Data Classification

The aim of the classification process is to verify if the suspended account produces more predictive contents for hate speech classification than the active accounts. This was implemented by training the classifier on the tweets from suspended/active accounts as hateful samples and testing the classifier on an unseen annotated data set. The Random Forest (RF) algorithm was adopted because it iteratively selects a random sub-sample of features in the training stage and trains multiple decision trees before predicting the outputs and results are averaged to maximize the reduction in classification error [19]. The Random Forest classifier was trained to contain 100 trees. Figure 1 illustrates our workflow used for investigation.

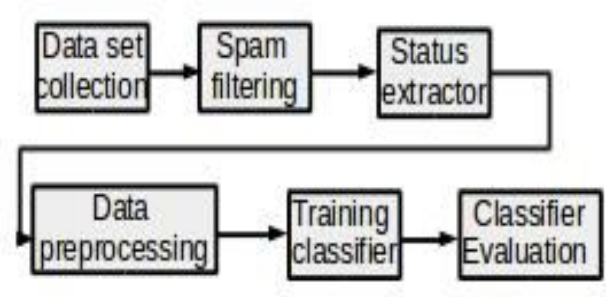

FIGURE 1. Examining the Suspended account text

5 Results and Discussion

In this section, we show our experimental results on emotion analysis to reflect the characteristics of tweets collected from different types of accounts. Also, the classification results are presented to compare the performance of the two classifiers trained on tweets collected from suspended and active accounts, respectively.

5.1 Datasets characterizations

In Figure 2 we show that the suspended accounts provided more disgusted and less joy tweets whereas the active accounts (see Figure 3) provided more anger and less surprise tweets. However, neutral accounts, which contain non-hateful tweets only, show more fear and less joy emotions, similar to the situation of suspended accounts. Figure 2 illustrates that the tweets from the suspended accounts contain more disgust, fear, negative and sadness emotions than the tweets from the active accounts. Furthermore, comparing the three types of accounts, the suspended accounts produce the highest frequency of the emotions in total. The variation of the emotion usage among the three types of accounts suggests that we could characterize a data set based on the status of the Twitter accounts. Interestingly, suspended accounts produce a larger number of tweets which contain all the ten emotions. This means that this sort of accounts tried to use positive emotions (e.g. trust emotion) along with the negative emotions to sooth the attitude presented in tweets. Finally, we could notice that the frequency of negative emotions is higher in the neutral tweets (see Figure 4) than in the hateful tweets from the active accounts, which indicate that the negativity does not always mean hate.

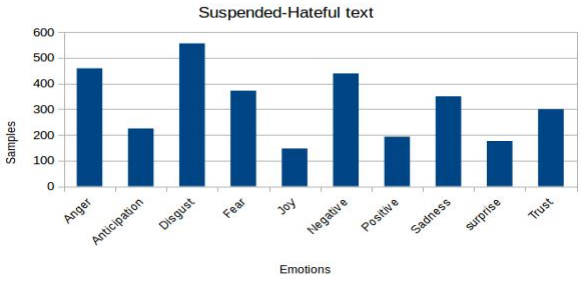

FIGURE 2. Suspended Accounts

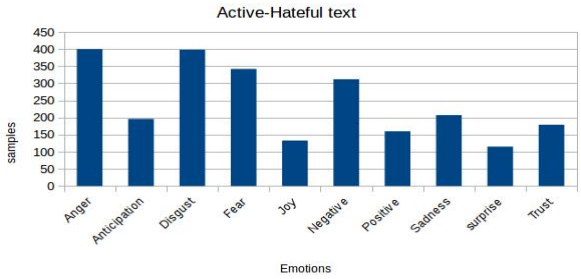

FIGURE 3. Active Accounts

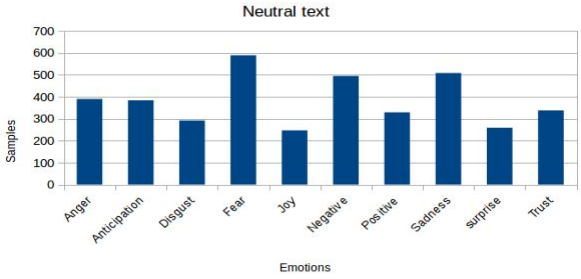

FIGURE 4. Neutral Accounts

5.2 Classification

Table 1 shows the results of testing the classifiers on tweets collected from different types of accounts, i.e. the results show comparison between the suspended account and the active account. The aim of this classification is to verify which type of accounts would provide tweets containing better features for accurate prediction of hateful and benign samples.

The results suggest that overall the most representative features for classifying cyber hate are extracted from the tweets produced by the suspended accounts. Use of tweets collected from suspended accounts results in higher performance than using the ones from active accounts by 16% in terms of average F-measure for both hateful and benign text. When the classifier was trained on tweets from suspended accounts for predicting both hateful and benign samples, the performance was improved in comparison with the case of training on tweets from active accounts. The results indicate that the users of suspended accounts produce more defined hateful terms than the users of active accounts. This suggests us to expect more subtle harassment to be produced by the suspended accounts.

TABLE 1. Classification Results by Testing the classifier on unseen data set

DS	**Active Accounts' tweets**			**Suspended Accounts' tweets**		
RF	P	R	F	P	R	F
	0.53 FP=1607	0.55 FN=2236	0.54	0.71 FP= 892	0.73 FN=1060	0.71

6 Conclusion

In this paper, we showed supporting evidence that the suspended accounts (accounts that were suspended by Twitter due to posting violent contents) could be a subtle and reliable source of hate speech text, specifically hate speech that harm others. We extracted the emotion characteristics for tweets from suspended, active and natural accounts, which showed that the three sources of tweets (suspended, active and neutral ones) reflected different emotions. As the tweets from suspended accounts presented the overall higher frequency of anger, fear, disgust, negative and sadness emotions than the tweets from active accounts, where both sources of tweets could be annotated as hateful instances. We argued that the tweets from suspended accounts would be a potential source for retrieval of hateful contents and this was evaluated by training a text classifier on each source (from suspended/active accounts) of tweets and then evaluating the performance of the two classifiers in terms of classifying unseen hateful tweets. The results showed that the use of tweets from suspended accounts could overcome the limitation of using tweets from active accounts, i.e. the rate of detecting the hateful tweets is increased by 16%.

For the problems (cost and reliability) related to the human annotation process, as mentioned previously, we proposed to leverage the text produced by suspended accounts, which are not spam, for augmenting hate speech data sets. We also proposed to collect tweets in the aftermath of hateful events and apply spam filtering to the data set. After that, we suggested to extract the tweets, which are posted from suspended accounts and are related to a specific event. This proposal will be mainly aimed at researchers who need big data sets for which the annotation process would be costly or unreliable.

Acknowledgements

The authors would like to acknowledge support for research reported in this paper from the Shaqra University.

References

[1] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter," in *Proceedings of NAACL-HLT*, 2016, pp. 88–93.

[2] B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki, "Measuring the reliability of hate speech annotations: The case of the european refugee crisis," *arXiv preprint arXiv:1701.08118*, 2017.

[3] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016, pp. 145–153.

[4] G. Xiang, B. Fan, L. Wang, J. Hong, and C. Rose, "Detecting offensive tweets via topical feature discovery over a large scale twitter corpus," in *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 2012, pp. 1980–1984.

[5] P. Burnap and M. L. Williams, "Us and them: identifying cyber hate on twitter across multiple protected characteristics," *EPJ Data Science*, vol. 5, no. 1, p. 11, 2016.

[6] ——, "Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making," *Policy & Internet*, vol. 7, no. 2, pp. 223–242, 2015.

[7] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," *arXiv preprint arXiv:1703.04009*, 2017.

[8] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 2, no. 3, p. 18, 2012.

[9] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, "Hate speech detection with comment embeddings," in *Proceedings of the 24th International Conference on World Wide Web*. ACM, 2015, pp. 29–30.

[10] S. Volkova and E. Bell, "Identifying effective signals to predict deleted and suspended accounts on twitter across languages." in *ICWSM*, 2017, pp. 290–298.

[11] ——, "Account deletion prediction on runet: A case study of suspicious twitter accounts active during the russian-ukrainian crisis," in *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, 2016, pp. 1–6.

[12] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Measuring# gamergate: A tale of hate, sexism, and bullying," in *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 2017, pp. 1285–1290.

[13] ——, "Mean birds: Detecting aggression and bullying on twitter," in *Proceedings of the 2017 ACM on Web Science Conference*. ACM, 2017, pp. 13–22.

[14] P.-C. Lin and P.-M. Huang, "A study of effective features for detecting long-surviving twitter spam accounts," in *Advanced Communication Technology (ICACT), 2013 15th International Conference on*. IEEE, 2013, pp. 841–846.

[15] R. Plutchik, *Emotions and life: Perspectives from psychology, biology, and evolution.* American Psychological Association, 2003.

[16] C. Strapparava, A. Valitutti *et al.*, "Wordnet affect: an affective extension of wordnet." in *Lrec*, vol. 4. Citeseer, 2004, pp. 1083–1086.

[17] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word–emotion association lexicon," *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, 2013.

[18] P. Sobhani, S. Mohammad, and S. Kiritchenko, "Detecting stance in tweets and analyzing its interaction with sentiment," in *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, 2016, pp. 159–169.

[19] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.