Classification of Hate and Non-Hate Tweets using Waseem Data

Harika Kanakam
Indiana University
hkanakam@iu.edu

Kamalesh Kumar Mandakolathur Guruprasad
Indiana University
kmandak@iu.edu

Abstract

This project aimed to replicate the experiment conducted by Waseem and Hovy (2016) on identifying abusive language on Twitter using machine learning techniques. We retrieved tweets from the past 2-3 months using the hashtag list of Waseem and Hovy and then annotated them to train and optimize a classifier for identifying abusive language. We categorized the tweets into two categories, sexism/racism, and none. We trained various machine learning models using the Waseem dataset and then tested it on the new data. Our results showed a lower accuracy than the Waseem dataset, which led us to conclude that there may be discrepancies in the hashtags of newly annotated tweets data. We identified a major possible cause of the discrepancy: MKR hashtag mostly refers to cryptocurrency tweets instead of referring to 'My Kitchen Rules'. In conclusion, our findings suggest that further research is needed to identify the most effective machine learning models and techniques for detecting abusive language on social media platforms.

1. Introduction

Hate speech on social media platforms, particularly Twitter, is a growing problem that poses significant challenges for online safety and freedom of expression. In response, researchers have turned to machine learning and natural language processing techniques to develop automated hate speech detection systems. Here, we examined a selection of papers that address the problem of detecting hate speech and offensive language on social media, particularly Twitter.

Machine learning techniques(Waseem and Hovy's (2016) study) are used to identify hateful users and words on Twitter. Their findings indicated that certain words and symbols can be predictive of hate speech, but they cautioned that the results could be affected by social biases. Similarly, Ribeiro et al. (2018) also used machine learning to identify hateful users on Twitter and found that hateful users tend to use more explicit and violent language than non-hateful users.

Kwok and Wang (2013) focused on detecting hate speech against a specific demographic, black people. They found that using a combination of lexicon-based and machine learning-based approaches could effectively identify hate speech tweets. Meanwhile, Chen et al. (2012) used machine learning to detect offensive language on social media to protect adolescent online safety and found that incorporating both linguistic and contextual features improved the performance of their detection system.

In addition, there are researches that differ in their approaches, techniques, and goals. For instance, Davidson et al. (2017) explored the problem of offensive language and hate speech detection and proposed a new dataset to evaluate the performance of automated systems. Sahoo et al. (2022) focused on detecting unintended social bias in toxic language datasets, highlighting the importance of fair and unbiased hate speech detection systems. Furthermore, Winter and Kern (2019) used Convolutional Neural Networks (CNNs) to perform multilingual hate speech detection on Twitter, while Pitsilis et al. (2018) used deep learning techniques to detect offensive language in tweets.

SemEval-2019 Task 6 OffensEval discusses the challenge which aimed to identify and categorize offensive language in social media. Several papers presented different approaches and techniques for hate speech detection, highlighting the complexity and importance of the problem, and the need for continued research to improve the accuracy and fairness of hate speech detection systems.

Overall, the importance of detecting and mitigating hate speech on social media platforms, particularly Twitter, is highlighted and a comprehensive overview of recent research in this area is provided. This emphasizes the variety of approaches and techniques employed in the studies, as well as their shared focus on the challenge of hate speech detection and the need for continued research to improve the accuracy and fairness of detection systems.

2. Data

We collected a total of 328 tweets over a period of 2 months using the public Twitter search API, filtering for non-English tweets. The resulting corpus consists of tweets containing both potentially offensive and non-offensive language, providing a realistic data set for studying hate speech on social media. We did not balance the data, as hate speech is a limited phenomenon, and our aim was to reflect real-world conditions as accurately as possible.

To ensure that we obtained a diverse range of tweets, we used the public Twitter search API to collect the entire corpus, filtering for tweets written in English. This corpus construction allowed us to collect non-offensive tweets that contained both clearly offensive words and potentially offensive words but remained non-offensive in their use and treatment of the words. For example, we found that even though the phrase "islam terrorism" is one of the most frequent in racist tweets, it can also occur in perfectly innocuous tweets, such as "If you are sincere in searching for the truth, God will lead you to it and then to heaven without effort #Islam #TheTruth #terrorism".

The list of hashtags that are considered to create the corpus are 'victimcard', 'BanIslam', 'FeminismIsCancer', 'MKR', 'whitegenocide', 'asian drive', 'SJW', 'islam terrorism', 'WomenAgainstFeminism', 'feminazi', 'immigrant', 'nigger'. These hashtags are used to retrieve the tweets from the past 2-3 months using the twitter API (https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api)

Later, we preprocessed the corpus and removed stop words, RT, screenname, and punctuation. We then manually annotated the corpus data, with a total of 328 tweets being annotated. Of these tweets, 91 were identified as containing sexist or racist content, while 237 were found to contain neither sexist nor racist content. Sexism/racism tweets are labeled as 1 and the tweets that are neither sexism nor racism are labeled as 2. It is important to note that since hate speech is a limited phenomenon, we did not balance the data, in order to provide a more realistic data set.

Detecting hate speech can be a challenging task, as it may not always be expressed through obvious racist or sexist slurs. Moreover, due to differences in exposure and knowledge of hate speech, identifying it can be subjective. A critical thought process is required to categorize the tweets. To provide a clear decision list for identifying hate speech, we used a set of criteria that is similar to Waseem and Hovy's (2016). These criteria negate the privileges observed in McIntosh (2003), where they occur as ways to highlight importance, ensure an audience, and are derived from common sense.

A tweet is considered abusive if it

1. uses a sexist or racial slur.
2. attacks a minority.
3. seeks to silence a minority.
4. criticizes a minority (without a well founded argument).
5. promotes, but does not directly use, hate speech or violent crime.
6. criticizes a minority and uses a straw man argument.
7. blatantly misrepresents truth or seeks to dis- tort views on a minority with unfounded claims.
8. shows support of problematic hash tags. E.g. "#BanIslam", "#whoriental", "#whitegenocide"
9. negatively stereotypes a minority.
10. defends xenophobia or sexism.
11. contains a screen name that is offensive, as per the previous criteria, the tweet is ambigu- ous (at best), and the tweet is on a topic that satisfies any of the above criteria.

During the annotation process of the collected tweets, a significant observation was made regarding the hashtags used in tweets related to #mkr. Contrary to what was expected, these tweets were mostly related to the cryptocurrency 'Maker' (MKR) and not the popular television show 'My Kitchen Rules' (MKR). On the other hand, tweets related to the hashtag #BanIslam were predominantly categorized as racist.

Another interesting finding during the annotation process was the low number of users who disclosed their location while tweeting. Only 1% of the total users revealed their location, which could be attributed to concerns over privacy or simply not wanting to share that information. It is important to note that the labeling of tweets as sexist or racist was not solely based on the use of specific terms or slurs, but also on the context and intent behind the tweets.

3. Experiments

We performed two baseline models, namely logistic regression and SVM, on the Waseem dataset to classify hate speech on social media. The performance of these models was then tested on a separate test dataset from Waseem, as well as on Twitter data.

Initially, We normalize the data similar to Waseem and Hovy's (2016) by removing stop words, with the exception of "not", special markers such as "RT" (Retweet) and screen names, and punctuation. we then trained the

Waseem data with logistic regression and SVM models and tested with Waseem data using scikit-learn package in python.

An accuracy of 0.82 with logistic regression and an accuracy of 0.83 with SVM is obtained. After experimenting with baseline models, SVM model is considered for further analysis. This suggests that the baseline models did not generalize well to the Twitter data, highlighting the differences in language use and context between the two datasets.

To address this issue, we fine-tuned the SVM model by optimizing their hyperparameters. In order to pick the most suitable features, we performed a grid search over all possible feature set combinations. The optimal hyperparameters obtained for SVM model are observed to be as below

```
{
    "clf_C": 0.1,
    "clfkernel": "linear",
    "vectmax_df": 0.5,
    "vect_ngram_range": (1, 2)
}
```

The best score achieved after fine-tuning for the Waseem test data was 0.85, and for the Twitter data, it was 0.73.

The results indicate that the fine-tuned SVM model slightly improved its performance on the Waseem test data compared to the baseline model. However, the accuracy on the newly annotated tweets data is comparatively lower than the accuracy obtained with Waseem test data after training with Waseem train data.

4. Results:

The accuracy of the newly annotated data is lesser than the accuracy of the Waseem test data. We generated classification reports from the predictions of the test data. A detailed overview of the classification results(including accuracy, F-1 score, precision and recall for both the categories) for both Waseem test data and newly annotated tweets data after fine-tuning with SVM model is shown below.

	precision	**recall**	**f1-score**	**support**
1	0.79	0.71	0.75	495
2	0.87	0.91	0.89	1076
accuracy			0.85	1571
macro avg	0.83	0.81	0.82	1571
weighted avg	0.85	0.85	0.85	1571

Table 1: Classification report for waseem test dataset

	precision	**recall**	**f1-score**	**support**
1	0.50	0.48	0.49	89
2	0.81	0.82	0.81	237
accuracy			0.73	326
macro avg	0.65	0.65	0.65	326
weighted avg	0.72	0.73	0.73	326

Table 2: Classification report for Newly annotated

The results of newly annotated tweets from the last 2-3 months are lower than those reported by Waseem and Hovy. This shows discrepancy in hashtag distribution and intensity of abuse.

From the results, it can be observed that calculated F1-score for Waseem test data is 0.75 for label 1(Hate Speech) and 0.89for label 2(Non Hate Speech). Whereas, the calculated F-1 score for newly annotated tweets data is 0.49 for label 1(Hate Speech) and 0.81 for label 2(Non Hate Speech).

And it can also be inferred that accuracy for Waseem test data is 0.85 and accuracy for newly annotated tweets data is 0.73. It is clear that both the F-1 score and accuracy for the tweets data is lower than the accuracy of the Waseem test data even after fine-tuning the SVM model.

5. Observation

With the lowered F-1 score and accuracy for the tweets data, we assume that there are discrepancies in the newly annotated tweets data. To investigate the discrepancies, we further analyzed the distribution of hashtags and calculated explicit abuse for each hashtag in the tweets.

5.1 Explicit Abuse

To evaluate explicit abuse, we considered both the raw data waseem dataset and newly annotated tweets before preprocessing. Using the considered data, the number of tweets with label 1 were filtered. The abusive words from “hate_lexicon_wiegand” text files are read and loaded to the dictionary. Each tweet in the test dataset was iterated and hashtags were extracted. If the abusive word is present in the tweet, the count for explicit abuse is incremented. This will obtain the total count of number of tweets which has explicit abusive words with hashtags used in Waseem data.

Finally, the percentage of tweets with explicit abuse for each hashtag in all the tweets is calculated and results are obtained. This percentage is calculated as ratio of number of tweets with label 1 and number of tweets with explicit abuse. For example, below table shows number and percentage of explicit abuse for hashtag **#mkr**

Data Set	**Number of tweets with label 1**	**Number of tweets with explicit abuse**	**% of label 1 tweets with explicit**
Waseem	418	52	8.04 %
Newly annotated tweets	5	0	0%

Table 3: Calculated explicit abuse percentage for MKR hashtag in both of the Waseem test data and newly annotated tweets data.

#mkr hashtag which had tweets flagged as sexist in waseem dataset, now has 0 tweets flagged as sexist in newly annotated tweets. Similarly, #immigrant which had lot of tweets flagged as racist in waseem dataset, has fewer tweets flagged as racist in newly annotated tweets.

Total percentage of explicit abuse is calculated as the ratio of the number of tweets which has explicit abusive words to the number of tweets that are labeled as 1 (racism/sexism). The results of the same calculated percentages is as mentioned in the table below

Data Set	**Total % of explicit abuse**
Waseem	31.51
Newly annotated tweets	27.30

Table 4: Calculated Percentage of explicit abuse in both Waseem test data and newly annotated tweets data for all hashtags.

From table 4, it can be observed that there is a reduction in the overall percentage of explicit abuse tweets in the newly annotated tweet dataset compared to Waseem test data.

5.2 Change in distribution

Waseem	**Newly annotated tweets**
#mkr #mkr2015 #islam #notsexist #killerblondes #isis #katandandre #iraq #syria #mkr	#feminismiscancer #whitegenocide #islam #immigrant #victimcard #terrorism #feminism #banislam #radfem #radfems

Table 5: Distribution of top 10 #hashtags

The above table shows a change in distribution of hashtags between waseem test data and newly annotated tweets.

	Label 1	**Label 2**
Mean	124.61	120.48
Std.	49.47	49.27
Min	31	250
Max	14	222

Table 6: Overview of lengths in characters, subtracting spaces

It can be inferred that the mean and standard deviation of the lengths in characters for the newly annotated tweets data is comparitively more than that of Waseem and Hovy's (2016). However, there is no significance difference in the standard deviations and overall length in characters between both the labels for the newly annotated data.

5.3 Newer hashtags

From table 5, distribution of top 10 hashtags it can be observed that newer hashtags like **“#radfem”**, **“#radfems”** were found in the tweets which are labeled as hate speech which are not mentioned in the Waseem and Hovy's (2016). This signifies change in trend in the use of hashtags for hate speech. Similarly, many new hashtags could have been used for hate speech which are not used in during the training phase of the models.

New Hashtags
#Ukraine
#istandwithrussia
#freetate
#istandwithputin
#NAFO
#Hindus
#jesuisparis
#blacklifematters
#war
#lovewins

Table 7: New Hashtags found in newly annotated tweets data

The above table shows newer hashtags which are present in newly annotated data but not in the previous waseem dataset.

We tried to further investigate the distribution of hashtag across geographic locations by obtaining the location of the users. As most of the users did not disclose the location details, this couldn’t be further investigated based on the geographic distribution.

6. Conclusion

We presented a list of criteria based on critical race theory to identify racist and sexist slurs on social media. Our study showed that using SVM and logistic regression models on the Waseem dataset resulted in relatively high accuracy scores for identifying hate speech. However, the model’s performance dropped when tested on Twitter data, suggesting that further efforts are needed to improve their generalization to different datasets.

We explored different approaches to address this issue, such as fine-tuning the models. We aimed to investigate the classification of demographic information, such as location, and calculated the percentage of explicit abuse to investigate the discrepancies in the newly annotated tweet data. We can further investigate the cause of discrepancies and analyze the trend of hate speech in order to generalize the model. While the problem of hate speech on social media remains challenging, our study provides a foundation for future research to build upon.

References

[1] Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. Proceedings of the NAACL Student Research Workshop, 88-93. https://aclanthology.org/N16-2013/

[2] Ribeiro, M., Calais, P., Santos, Y., Almeida, V., & Meira Jr., W. (2018). Characterizing and Detecting Hateful Users on Twitter. Proceedings of the International AAAI Conference on Web and Social Media, 12(1). https://ojs.aaai.org/index.php/ICWSM/article/view/15057.

[3] Kwok, I., & Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1781-1791. https://ojs.aaai.org/index.php/AAAI/article/view/8539

[4] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012a. Detecting offensive language in social media to protect adolescent online safety. In Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom), pages 71–80. IEEE, September. https://www.cse.psu.edu/~sxz16/papers/SocialCom2012.pdf

[5] The Effects of User Features on Twitter Hate Speech Detection by Elise Fehn Unsvag and Bjorn Gamback. https://aclanthology.org/W18-5110.pdf

[6] Automated Hate Speech Detection and the Problem of Offensive Language” by Thomas Davidson et al. presented at ACL 2017. https://ojs.aaai.org/index.php/ICWSM/article/view/14955/14805

[7] "Detecting Unintended Social Bias in Toxic Language Datasets” by Nihar Sahoo, Himanshu Gupta, Pushpak Bhattacharyya. https://arxiv.org/pdf/2210.11762.pdf

[8] “Know-Center at SemEval-2019 Task 5: Multilingual Hate Speech Detection on Twitter using CNNs” by Kevin Winter, Roman Kern. https://aclanthology.org/S19-2076.pdf

[9] “Detecting Offensive Language in Tweets Using Deep Learning” by Georgios K. Pitsilis, Heri Ramampiaro and Helge Langseth. https://arxiv.org/pdf/1801.04433.pdf

[10] “SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)” by Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, Ritesh Kumar. https://aclanthology.org/S19-2010.pdf