Racism Detection by Analyzing Opinions Through Sentiment Analysis of Tweets from Different Immigration Crisis.

Hiba Lubbad

MSc in Data Science
The University of Bath
2022

Racism Detection by Analyzing Opinions Through Sentiment Analysis of Tweets from Different Immigration Crisis.

Submitted by: Hiba Lubbad

Copyright

Declaration

This dissertation is submitted to the University of Bath in accordance with the requirements of the degree of [INSERT YOUR COURSE TITLE HERE] in the Department of Computer Science. No portion of the work in this dissertation has been submitted in support of an application for any other degree or qualification of this or any other university or institution of learning. Except where specifically acknowledged, it is the work of the author.

Racism Detection by Analyzing Opinions Through Sentiment Analysis of Tweets from Different Immigration Crisis.

Abstract

<The abstract should appear here. An abstract is a short paragraph describing the aims of the project, what was achieved and what contributions it has made.>

Racism Detection by Analyzing Opinions Through Sentiment Analysis of Tweets from Different Immigration Crisis.

Contents

Acknowledgements

Add any acknowledgements here.

Racism Detection by Analyzing Opinions Through Sentiment Analysis of Tweets from Different Immigration Crisis.

Introduction

1.1 Background and Description

Twitter has become a dominating socio-political platform allowing live news updates that outpace news outlets by allowing witnesses and participants to share the events on the platform quickly. Twitter also allows people worldwide to voice their opinions on the current crisis. Amongst these voices, misuse of the platform has seen emerging racism and discrimination in various forms. This ranges from memes to openly racist remarks that incite social instability and violence. The recent Ukraine immigration crisis shows that Twitter users have used the platform to voice their opinions, but it has seen various racist tweets from individuals to media outlets. Online Social Networks (OSNs) have noticeably played a significant role in such crises and have proven to be a reliable and fruitful source of data for analysis and studying political discourse. Social media being the leading source of racism, opinions dissemination should be monitored, and racist remarks should be detected and blocked timely. Sentiment analysis provides a powerful tool to mine such data to analyze emotions and opinions. Moreover, the different type of discussions on racism illustrates geographical variability in racial attitudes and sentiment. Therefore, analyzing how people, events, and circumstances are represented provides insights into user communication and problems surrounding racism can be investigated.

Automatic hate speech detection using machine learning algorithms is still new and requires extensive research efforts from both industry and Academia. It has been estimated that on a yearly basis there are hundreds of millions of euros are being invested to combat hate speech (Zhang et al. 2020). While social media platforms have implemented strict integrity and hateful contact policies, the problem remains complex because it involves several layers of complexity: computational complexity, given the large volume of content, as well as the subtleties and cultural aspects of each language, the problem of low resource languages and natural language's intrinsic ambiguity (Goutsos et al.,2021). This research aims to perform sentiment analysis on tweets from the Afghanistan and Ukraine crises and then implement various learning models to detect racism in

tweets. The research will target tweets that contain hateful and racist speech aimed at refugees and migrants. Therefore, it will try to improve on existing models throughout this paper.

1.2 Dataset

A Dataset for the Russo-Ukrainian Crisis has been extracted by researchers since the 1st week of the crisis and is being updated daily and publicly available on GitHub (Haq et al., 2022). However, due to twitter guidelines the repository only contains twitter IDs, but the tweets can be extracted using open-source tools such as Twarc, Tweepy, or Hydrator. The Afghanistan dataset will be scraped similarly. Twitter's API provides access to twitter objects that contain "a long list of 'root-level' attributes, including fundamental attributes such as *id*, *created_at*, and *text"* (Twitter).

```
Index(['id', 'conversation_id', 'created_at', 'date', 'time', 'timezone',
       'user_id', 'username', 'name', 'place', 'tweet', 'language', 'mentions',
       'urls', 'photos', 'replies_count', 'retweets_count', 'likes_count',
       'hashtags', 'cashtags', 'link', 'retweet', 'quote_url', 'video',
       'thumbnail', 'near', 'geo', 'source', 'user_rt_id', 'user_rt',
       'retweet_id', 'reply_to', 'retweet_date', 'translate', 'trans_src',
       'trans_dest'],
      dtype='object')
```

Figure 1. Meta Data Available in a Tweet

1.3 Methods

According to previous research, the analysis could be done through exploring and combining both text and image modalities as opposed to the traditional text-only (Goutsos). Specifically, users often use messages encoded in images to avoid NLP-based hate speech detection systems.

An ensemble model will be proposed that makes use of recurrent neural networks. For performance comparison, several well-known machine learning models are implemented using the optimized parameters such as decision tree (DT), random forest (RF), logistic regression (LR), k nearest neighbour (KNN), and support vector machines (SVM). Term frequency-inverse document frequency (TF-IDF) and a bag of words (BoW) are studied as feature extraction techniques (Lee, E).

Racism Detection by Analyzing Opinions Through Sentiment Analysis of Tweets from Different Immigration Crisis.

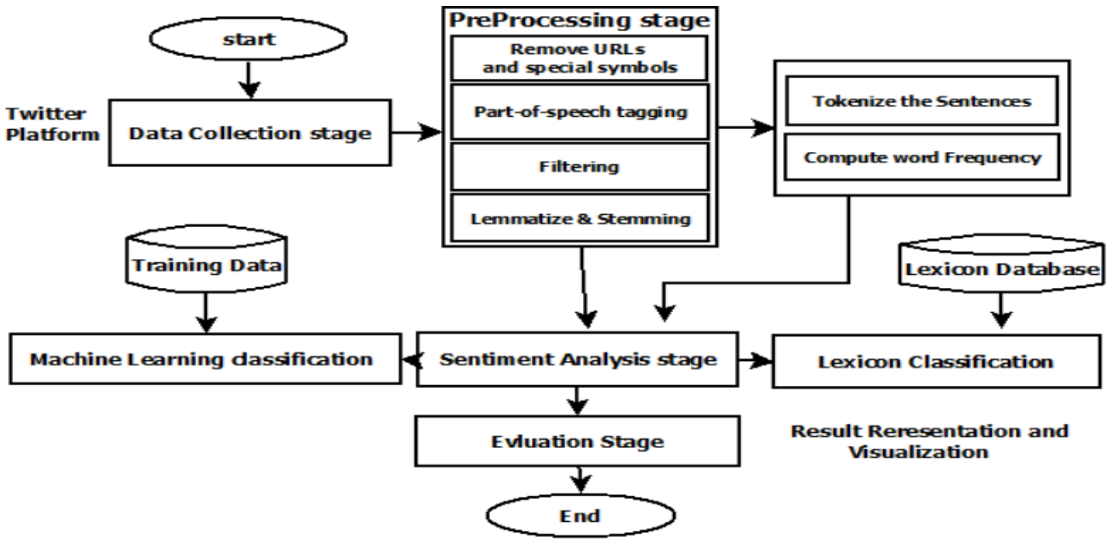
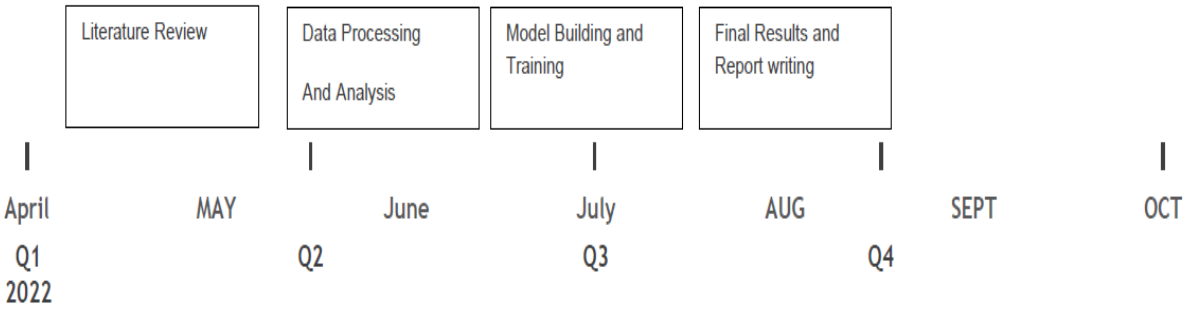


Figure 2. Approach follow diagram of susing NLP, text mining, and Sentiment Analysis techniques.

1.4 Gant Chart



Literature and Technology Survey

As more users join social media, the spread of hate speech has increased, thus pushing for urgent intervention from governments, enterprises, and academia. Therefore, there has been a rise in emerging research and publications on hate speech detection in different issues and languages. This paper will focus mainly on hate speech detection efforts and publications of anti-immigration speech.

A. General Hate Speech Detection

Parihar et al. (2021) provide a discussion on various relevant research in the field of hate speech detection. The paper generalises the structure of hate speech (Fig 3.) and argues that most of the publications utilise social media-based datasets with a different variety of hate speech labelling. For example, Malmalsi et al. (2017) annotated the dataset into three different labels: hate, offensive, and OK. Meanwhile, Cao et al. (2020) proposed a DeepHate architecture that classified tweets as either inappropriate or normal, achieving precision as high as 92.48%. The next step in building the hate speech detection model is pre-processing and feature extraction. Zhang, Robinson, and Tepper (2018) categorise feature extraction methods into classical and deep learning methods. The classical methods are manual feature engineering often used by machine learning algorithms such as Naive Bayes, Logistic Regression, and SVMs. The most commonly used and shown to be highly effective classical method is the surface features such as “bag of words, word and character n-grams” (p.747). N-grams are defined as a sequence of consecutive words. Waseem and Hovy (2016) provide findings that n-grams are an effective option for hate speech detection by analysing the impact of using several extra-linguistic features in conjunction with character n-grams for hate speech detection. Another common method uses sentiment analysis to detect the polarity expressed in a tweet. For example, Jiang and Suzuki (2019) used sentiment analysis to detect hate speech from tweets.

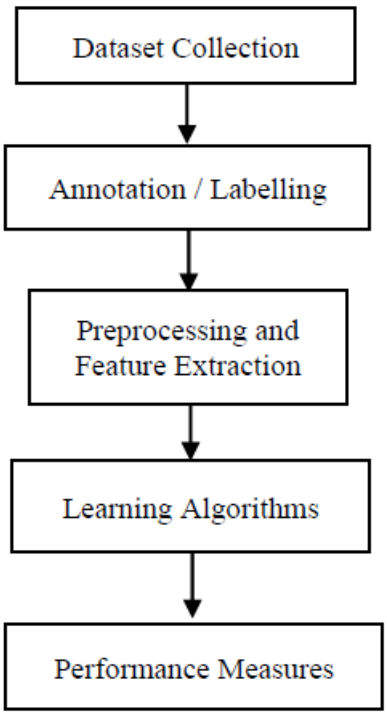


Figure 3. General structure of hate-speech detection model

Moreover, Gitari et al. (2015) present a lexicon-based approach for hate speech detection where they use subjectivity and semantic features to create a lexicon used by a hate speech detection classifier. However, the proposed work of Cao et al. has shown that using multifaceted representations of the text such as “word embeddings, sentiments, and topical information” outperforms methods relying on single textual features. The next essential step in building a hate speech detection model is to choose the appropriate classification algorithm. Waseem and Hovy's work () served as a foundation for several other studies that looked at predictive factors for hate speech detection. They provided access to their massive 16K twitter corpus, which is dedicated to hate speech detection in the English language. Many publications have discussed the literature on the different algorithms used in hate speech detection such as Putri et al, Ketsbaia et al., and Hassan et al. Initially machine learning algorithms can be divided into supervised, semi-supervised, and unsupervised. Much of the research on hate speech detection has been supervised, requiring manual labelling of the dataset. While the most popular classification algorithm is Support Vector Machines (SVM), others were used such as Naïve Bayes, Random Forests, Logistic Regression, and Decision Trees. There is no evidence that one algorithm generally outperforms the rest as the choice of

algorithm depends on the task and features extracted. For instance, Burnap and Williams (2014) tested different supervised algorithms for hate speech detection, and all the classifiers performed the same but varied accuracies when different features were tested.

Furthermore, Putri et al. conclude that Multinomial Naïve Bayes is the model for hate speech detection as it outperforms the others with an accuracy of 71%; while Chen et al. (2012) found that SVM is more accurate than Naïve Bayes at detecting offensiveness. As the study of HS detection has been a complex task, some research has attempted a different approach, such as transfer learning. For example, Rizoiu et al. have used transfer learning from an LSTM network coupled with ELMo embeddings and found that the model trained on multiple generic abusive language datasets will produce more robust predictions. Moreover, another approach done by Swamy et al. (2019) and Mozafari et al. (2019) is applying a Bidirectional Encoder Representations from Transformers (BERT) based transfer learning approach. The findings demonstrate the model's capacity to detect some biases in the collection or annotation of datasets. However, the biggest challenge in classifying abusive language across domains is dataset limitations and biases. Another gap remaining in hate speech detection research is the lack of research in languages aside from English. Nonetheless, Ranasinghe et al. used a social media dataset with multiple languages like English, German and Hindi and proposed a BERT-based architecture. There has also been very little research done on hate speech detection in Arabic (Aljarah et al., 2021).

B. Deep Neural Network Methods

More recently, hate speech detection has shifted towards exploiting deep neural network methods such as LSTMs, GRU, and CNNs combined with word embedding models such as ELMo and word2vec. Word embedding alleviated the data sparsity problem by introducing an additional semantic feature as it constructed distributed representations that introduce word dependency (Al-Hassan and Al-Dossari, 2019). One of the first attempts of Deep artificial Neural Networks in HS detection used paragraph2vec embeddings in a two-step classifier (Djuric et al., 2015). Moreover, according to Lilleberg et al. (2015), word2vec has piqued the curiosity of experts in the field and has performed well in detecting prejudice in social media posts, as shown by Yuan et al. which achieved an overall accuracy of 0.91. On the other hand, Pitsilis et al. (2018) employed an RNN model with word frequency vectorization instead of word embedding to build features for hate speech detection,

breaking the barrier of language reliance in the word embedding technique. For hate speech identification, their results exceeded existing state-of-the-art deep learning algorithms.

The use of deep learning has become more popular due to the advancement in graphic processing units and their ability to detect and extract features automatically. Deep learning methods have generally outperformed classical machine learning methods. For example, Zhang and Luo (2018) compared a deep neural network model to several classical machine learning methods and found by comparing the f1 scores that the neural network performed the best. Moreover, Ali, El Hammid, and Youssif's (2019) research compared the results of different deep neural network methods to SVM and Naïve Bayes for a classification sentiment analysis task and found that DNN methods outperformed greatly. However, this does not offer a golden rule where DNN methods always outperform classical methods as it greatly depends on the complexity of the hidden layers and the correct choice of algorithm and feature representation technique. Al-Smadi et al. (2018) work to support this statement as it compares the performance of RNNs and SVMs and shows that for a specific set of features the SVM greatly outperforms the RNN. Based on the literature, CNNs are well-known for acting as 'feature extractors' whereas RNN is well-suited to modelling ordered sequence learning issues. In the context of hate speech categorization, RNN learns word or character dependencies in tweets, whereas CNN extracts word or character combinations such as phrases and n-grams (Zhang Et al., 2018).

The most powerful combination that has been recently investigated of CNN+RNN is in theory established to be a powerful structure for capturing order information between CNN features. This combination has also performed well in practice on different tasks such as activity recognition. Empirical research has also shown that building a DNN with (GRU) instead of LSTM achieved comparably better results. This is due to their simpler structure thus making them easier to train and generalize on smaller datasets. Due to improvement in deep learning's performance, researchers have used a stacked ensemble deep learning model which is created by merging gated recurrent units (GRU), convolutional neural networks (CNN), and recurrent neural networks RNN. This is referred to as Gated Convolutional Recurrent-Neural Networks (GCR-NN). CNN extracts crucial characteristics for RNNs to produce good predictions, whereas GRU is at the top of the GCR-NN model for extracting significant features from raw text. For example, Alshalan and H. Al-Khalifa (2020) use several deep learning algorithms to explore the issue of hate speech on Saudi Twitter. BERT, CNN, GRU, and the ensemble of CNN and GRU (CNN+GRU) are used in a series of tests on two datasets. The CNN

model scores an F1 score of 0.79. Moreover, Lee et al. (2022) purpose a GCR-NN structure (Fig 4.) that was able to detect 97% of the tweets containing racist comments. In relation to research focused on racism and refugees, Zhang et al. (2018) presented a deep neural network that integrated convolutional and gated recurrent networks targeting hate speech detection that achieved very promising results.

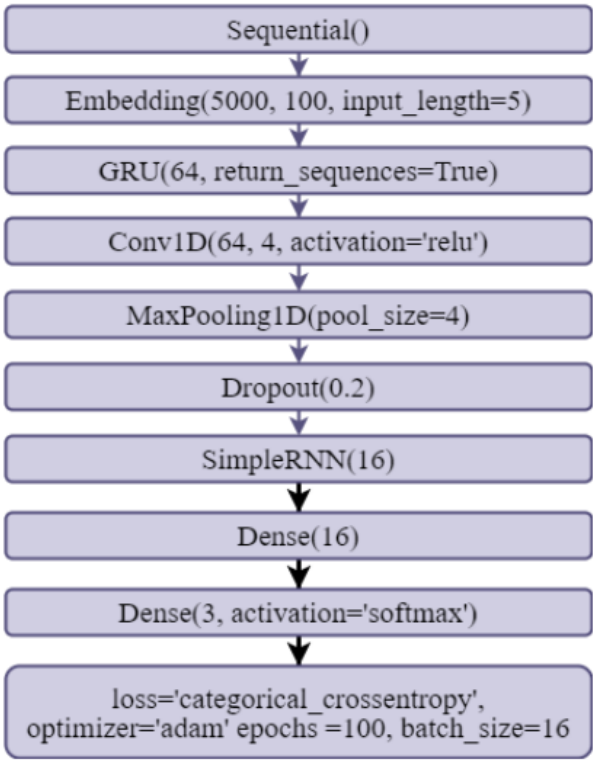


Figure 4. Strucutre of propose GRNN

This work uses the deep learning ensemble model to detect racist remarks on Twitter, taking into consideration the reported findings from deep learning models. By layering recurrent neural networks, the researchers hope to achieve high classification accuracy.

Bibliography

Al-Hassan, A. and Al-Dossari, H., 2019, February. Detection of hate speech in social networks: a survey on multilingual corpus. In *6th international conference on computer science and information technology* (Vol. 10).

Ali, N.M., Abd El Hamid, M.M. and Youssif, A., 2019. Sentiment analysis for movies reviews dataset using deep learning models. *International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol*, *9*.

Cao, R., Lee, R.K.W. and Hoang, T.A., 2020, July. DeepHate: Hate speech detection via multi-faceted text representations. In *12th ACM conference on web science* (pp. 11-20).

Chiril, P., Pamungkas, E.W., Benamara, F., Moriceau, V. and Patti, V., 2022. Emotionally informed hate speech detection: a multi-target perspective. *Cognitive Computation*, *14*(1), pp.322-352.

Djuric, N., Zhou, J., Morris, R., et al.: Hate speech detection with comment embeddings. In: Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Companion, pp. 29–30. ACM, New York (2015). https://doi.org/10.1145/2740908.2742760
G. K. Pitsilis, H. Ramampiaro, and H. Langseth, “Effective hate-speech detection in Twitter data using recurrent neural networks,” Appl. Intell., vol. 48, no. 12, pp. 4730–4742, Dec. 2018.

Gitari, N.D., Zuping, Z., Damien, H. and Long, J., 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, *10*(4), pp.215-230.

Haq, E.U., Tyson, G., Lee, L.H., Braud, T. and Hui, P., 2022. Twitter dataset for 2022 russo-ukrainian crisis. *arXiv preprint arXiv:2203.02955*.

J. Lilleberg, Y. Zhu, and Y. Zhang, “Support vector machines and Word2vec for text classification with semantic features,” Proc. 2015 IEEE 14th Int. Conf. Cogn. Informatics Cogn. Comput. ICCI*CC 2015, pp. 136–140, 2015.

Ketsbaia, L., Issac, B. and Chen, X., 2020, December. Detection of Hate Tweets using Machine Learning and Deep Learning. In *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)* (pp. 751-758). IEEE.

L. Jiang and Y. Suzuki, "Detecting hate speech from tweets for sentiment analysis," 2019 6th International Conference on Systems and Informatics (ICSAI), 2019, pp. 671-676, doi: 10.1109/ICSAI48974.2019.9010578.

M. Al-Smadi, O. Qawasmeh, M. Al-Ayyoub, Y. Jararweh, and B. Gupta, “Deep Recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels’ reviews,” J. Comput. Sci., vol. 27, pp. 386–393, 2018.

Mozafari, M., Farahbakhsh, R. and Crespi, N., 2019, December. A BERT-based transfer learning approach for hate speech detection in online social media. In *International Conference on Complex Networks and Their Applications* (pp. 928-940). Springer, Cham.

Parihar, A.S., Thapa, S. and Mishra, S., 2021, June. Hate Speech Detection Using Natural Language Processing: Applications and Challenges. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 1302-1308). IEEE.

Pitsilis GK, Ramampiaro H and Langseth H. Effective hate-speech detection in Twitter data using recurrent neural networks.Appl Intell 2018; 48(12):4730–4742.

Putri, T.T.A., Sriadhi, S., Sari, R.D., Rahmadani, R. and Hutahaean, H.D., 2020, April. A comparison of classification algorithms for hate speech detection. In *Iop conference series: Materials science and engineering* (Vol. 830, No. 3, p. 032006). IOP Publishing.

R. Alshalan and H. Al-Khalifa, ‘‘A deep learning approach for automatic hate speech detection in the Saudi Twittersphere,’’ Appl. Sci., vol. 10, no. 23, p. 8614, Dec. 2020.

Racism Detection by Analyzing Opinions Through Sentiment Analysis of Tweets from Different Immigration Crisis.

Rizoiu M, Wang T, Ferraro G, and Suominen H. Transfer Learning for Hate Speech Detection in Social Media. CoRR, abs/ 1906.

S. Malmasi and M. Zampieri, "Detecting hate speech in social media,"arXiv preprint arXiv:1712.06427, 2017.

Waseem, Z. and Hovy, D., 2016, June. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop* (pp. 88-93).

Zhang, Z., Robinson, D. and Tepper, J., 2018, June. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference* (pp. 745-760). Springer, Cham.

Zhang Z and Luo L. Hate speech detection: a solved problem? The challenging case of long tail on Twitter. arXiv preprint: arXiv:1803.03662, 2018.

Lee, E., Rustam, F., Washington, P.B., El Barakaz, F., Aljedaani, W. and Ashraf, I., 2022. Racism Detection by Analyzing Differential Opinions Through Sentiment Analysis of Tweets using Stacked Ensemble GCR-NN Model. *IEEE Access*.

Goutsos, D. (2021). Multimodal hate speech detection in greek social media. Multimodal Technologies and Interaction, 5(7), 34. doi:http://dx.doi.org/10.3390/mti5070034

Twitter. (n.d.). Tweet object | docs | twitter developer platform. Twitter. Retrieved April 5, 2022, from https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet

Racism Detection by Analyzing Opinions Through Sentiment Analysis of Tweets from Different Immigration Crisis.