# Modeling Hate Speech Detection in Social Media Interactions Using Bert

Gibran Mwadime[1], Moses Odeo[2], Boniface Ngari[3], Stephen Mutuvi[4]
[1,2,3,4]*Department of Computer Science, Multimedia University of Kenya*

***Abstract*—Hate speech propagation in social media sites has been happening over time and there is need to accurately identify and counter it so that those offended can seek redress and offenders can be punished for perpetrating the vice. In this paper, we demonstrate how fine tuning a pre-trained Google Bidirectional Encoder Representation from Transformers (BERT) model has been used to achieve an improvement in accuracy of classification of tweets as either hate speech or not. Random forests and logistic regression algorithms have been used to build baseline models with a publicly available twitter dataset from *hatebase.org.* To validate the BERT model, we collected data using Tweepy API and combined with data from *hatebase.org* for training. The results obtained show an improvement in accuracy of tweets classification as either hate speech or not from the baseline models by 7.22%.**

***Keywords:*** **Sentiment analysis, hate speech, social media, model, data**

## I. INTRODUCTION

Hate speech is defined as any form of expression that seeks or expresses hatred against a person or group of persons because of something they are associated with [6]. Social media has become an environment where people from all walks of life and from different geographical locations converge to share experiences, opinions and ideas. In the recent years, there has been a huge growth in the use of social media and misuse to propagate hate speech and related activities [7]. A lot of information is generated that contains ambiguities and a lot of noise making it difficult to decipher what it means. This calls for a reliable and accurate sentiment analysis tool to ensure text is understood clearly and categorized to the class it belongs appropriately. Differences in opinions have led to abuse and exchanges that result in hatred between parties involved. In the UK, moments after Butt used a van to run over pedestrians, twitter was abuzz with over 18 million tweets published within an hour containing happy messages, hatred towards Islam and support for violence [2].

Social media sites such as twitter and Facebook have been under intense pressure to monitor and control content posted in their platforms however it has not been easy due to intensiveness and time it takes to manually go through each post [10],[3]. Similarly it is problematic and difficult to censor user's posts because of freedom of speech. There is need to have an automated tool to accurately detect and classify tweets as either hate speech or not. It is for this reason that Natural Language Processing (NLP) techniques using machine learning come in handy to assist in analyzing tweets to identify hate content and report [8]. Sentiment analysis is performed to understand the tweets, draw correlations and classify them appropriately.

Our work makes three contributions to Natural Language Processing task of text classification. First we build baseline models using Logistic Regression and Random Forests algorithms with publicly available dataset from *hatebase.org.* A comparative evaluation of the performance of the two models is done and results analyzed. Secondly we collect data from twitter using defined keywords related to hate content creating a dataset that we use to validate our proposed model. We lastly build our proposed model implementing Transfer learning. A Bidirectional Encoder Representation for Transformers model using a Google pre-trained model was developed.

The remainder of this paper is structured as follows; in section II we discuss related work on hate speech classification in Twitter. In section III we describe the approach we used in our work. In section IV we analyze and discuss results for the models built. Recommendations and Conclusions are highlighted in Section V and VI respectively.

## II. RELATED WORK

A lot of work on detection of hate speech in twitter has been done and more work is still required to improve the accuracy metrics. The accuracy metrics are Precision, Recall and F_score. Lexical methods are effective in identifying offensive words but are ineffective in identifying hate speech [9].

Bag of words approaches usually tend to have high recall although they lead to high rates of false positives [1],[4]. This is because the presence of offensive words can lead to misclassification of tweets as hate speech.

[3] used four Convolutionary Neural Networks models to classify twitter hate speech data that was trained on character 4-grams, word vectors based on semantic information built using word2vec, randomly generated word vectors, and word vectors combined with character n-grams. The models were tested with 10 fold validations and resulted in higher precision of 85.66%, recall of 72.14% and an F score of 78.3%. The same dataset classified using Logistic Regression model by [10] had a precision of 72.87%, recall of

77.75% and F score of 73.89%.The character n grams contributed greatly to the higher precision levels.

Recently contextual word embedding models have been developed and have shown significant improvement in solving NLP tasks including text classification, question answering and semantic role labeling [5]. An embedding is created and is used as a feature representation in a classifier. They help in word sense disambiguation and addressing the problem of polysemy where a word has more than one meaning [5].

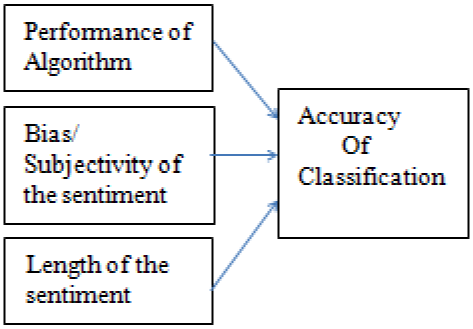
Performance of Algorithm
Bias/ Subjectivity of the sentiment
Length of the sentiment
Accuracy Of Classification

Fig.1 Conceptual framework for hatespeech detection

## III. RESEARCH DESIGN

This research used a publicly available dataset from *hatebase.org* and data collected using Tweepy API.Twitter was chosen because it is the most frequently used social media network generating a lot of data within a short period of time. We adopted a quasi-experimental design where experiments were carried out using the available datasets to model accuracy in classification of tweets as either hate speech or not.

We implement our research with three machine learning algorithms used for text classification; Logistic Regression and Random Forests and BERT. We build baseline models using Logistic Regression and Random Forests algorithms on annotated hate speech dataset from *hatebase.org* containing 24803 tweets. We train each model with data split into 70% train set containing 17348 tweets and 30% test set containing 7435 tweets and the results recorded. The dataset from *hatebase.org* is publicly available and has been used by other researchers previously in detection of hate speech in twitter. We pre-process the data by removing retweets, lower casing the tweets, removing punctuations, URL's, stop words and mentions. Hyper parameters are varied during training for each model so that we obtain the best result. The data collected using the API was used for validation of the BERT model after training it with the publicly available dataset.

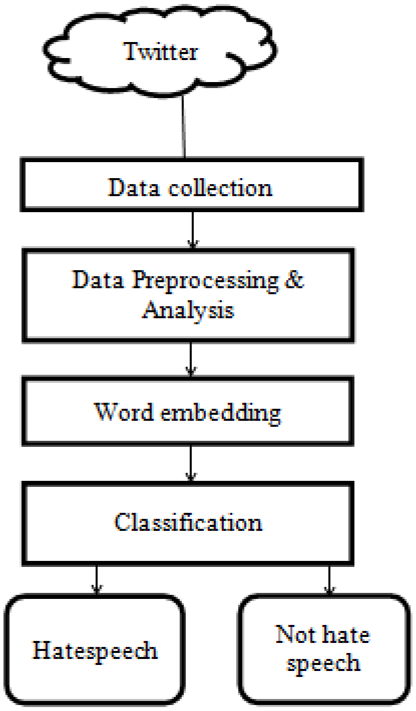

Fig.2 Modeling process

### *A. Research site*

Twitter was chosen because it is the most frequently used social media network generating a lot of data within a short period of time. Tweets are short up to a maximum of 140 words and have unique lexical and semantic features that are different from other types of text. Similarly twitter has an open platform where researchers can access and collect large amounts of data without much difficulty using API's unlike other social media sites like Facebook.

### *B. Population & Sample size*

For this research, a large amount of data was required for modeling and classification. The target population was a large dataset. We used a sample size of 25,000 tweets to build our model.

## IV. EXPERIMENTS AND DISCUSSION OF RESULTS

In this section, we seek to apply NLP task of classifying text as either hate speech or not using data collected from Twitter. We attempted to achieve the following research objectives;

1) To investigate the performance of Logistic Regression and Random Forests algorithms in hate speech classification.
2) To mine data from twitter for hate speech classification.
3) To develop a model for accurate hate speech classification.

### *A. Logistic Regression*

We developed the first baseline model using logistic regression with a hate speech dataset obtained from *hatebase.org*.

The dataset has 24803 tweets which we split into two sets; 70% train and 30% test set. The training set was composed of 17348 tweets while the test set was composed of 7435 tweets. The results for the model are as shown in Table I.

Table I Logistic Regression Model

| X_train 17348 | Maximum features | Random state | Accuracy (%) |
|---|---|---|---|
| X_test 7435 | 1500 | 5 | 80.65 |
| X_test 7435 | 1500 | 7 | 84.14 |
| X_test 7435 | 1500 | 10 | 84.30 |
| X_test 7435 | 1500 | 50 | 84.92 |
| X_test 7435 | 1500 | 52 | 84.34 |
| X_test 7435 | 1500 | 100 | 85.08 |

The classification report of the model is as shown below.

```
        === Classification Report ===

                Precision Recall F_score

  0              0.85      1.00    0.92
  1              0.00      0.00    0.00
  2              0.72      0.89    0.79
  3              0.00      0.00    0.00
  4              0.00      0.00    0.00
  5              0.00      0.00    0.00
   6             0.00      0.00    0.00
   7             0.00      0.00    0.00
   Micro avg    0.84      0.84    0.84
Macro avg      0.20      0.24    0.21
 Weighted avg  0.71      0.84    0.77
```

### *B. Random Forests model*

We developed a second baseline model using Random forests using the same dataset from *hatebase.org* and the results were recorded as shown in Table II.

Table II Random Forests Model

| X_train 17348 | Estimators | Random epochs | Accuracy (%) |
|---|---|---|---|
| X_test 7435 | 100 | 252 | 80.65 |
| X_test 7435 | 100 | 152 | 80.41 |
| X_test 7435 | 100 | 52 | 80.37 |

The classification report of the model is as shown below.

```
    === Classification Report ===

      Precision   Recall   F_score

    0    0.84      0.97     0.90
    1     0.31      0.08     0.12
    2     0.31      0.18     0.23
    3     0.29      0.07     0.11
    4     0.00      0.00     0.00
    5     0.00      0.00     0.00
    6     0.00      0.00     0.00
Micro avg    0.80     0.80     0.80
Macro avg    0.25     0.19     0.19
Weighted avg  0.73     0.80     0.75
```

### *C. BERT model*

We fine-tuned a BERT model with the dataset from *hatebase.org* used for training and data collected using Tweepy for validation of our model. The model applied fine tuning approach with a Google BERT pre-trained model repository from github cloned in our machine.

The results of the model are as shown in Table III below.

Table III BERT model

| Model | Bert |
|---|---|
| Accuracy | 91.36% |
| Precision | 0.75 |
| Recall | 0.87 |
| F_score | 0.81 |

### *D. Discussion of Results*

Logistic regression performs better in accuracy with 84.14% against 80.39%, Recall of 0.84 against 0.80, F_score of 0.77 against 0.77 while Random forests performs better on precision with 0.73 against 0.71.

It was noted that varying hyper parameters during training of the model affected the output result up to a point where further variations did not affect the results. Increasing the random epochs above 252 in Random Forests model did not increase the accuracy but remained constant.

The Bert model performed better with an accuracy of 91.36% proving to be a reliable model for automated hate speech detection. An F_score measure of 0.81 was recorded against 0.77 and 0.75 for Logistic Regression and Random Forests model signifying a great improvement in the model's performance.

The training of Bert model is a computer resource intensive task that was observed during experiments.

## V. RECOMMENDATIONS

The fine tuning approach was applied on an existing Google BERT pre-trained model for our experiments to achieve 91.36% accuracy and F_Score of 0.81.We recommend further research that will involve fine tuning of the BERT model to investigate variation in the performance evaluation metrics. With further fine tuning, Precision and Recall measures may be recorded as they were not obtained in our experiments. This exercise however requires a lot of memory space to carry out the experiments.

## VI. CONCLUSION

The research achieved the objectives set in the beginning. A comparison of performance of Logistic Regression and Random Forests algorithms used for Natural language processing task of text classification was carried out and Logistic Regression was found to perform better than Random Forests.

Using Tweepy API, we were able to collect user generated data from twitter. We preprocessed the data and were used as a validation set for the model developed.

A BERT model that performed better than the baseline models was developed and the results obtained were quite impressive recording an accuracy improvement of 7.22%. BERT utilizes transfer learning principle in its implementation and from the results; it is evident that transfer learning contributes greatly to the improvement of results in any learning task.

## REFERENCES

[1] Burnap, P., and Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet* 7(2):223–242.

[2] Esteve, M., Miró-Llinares, F., & Rabasa, A.(2018). Classification of tweets with a mixed method based on pragmatic content and meta-information. *International Journal of Design & Nature and Ecodynamics, 13*(1), 60– 70.

[3] Gamback, B. and Sikdar, U. K. (2017). Using convolutional neural networks to classify hate speech. *In Proceedings of The First Workshop on Abusive Language Online*, pages 85–90. Association for Computational Linguistics.

[4] Kwok, I., and Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. In *AAAI*.

[5] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. (2018). Deep contextualized word representations. *In NAACL*.

[6] Mendel, T., Herz, M., and Molnar, P. (2012). Does international law provide for consistent rules on hatespeech? *The content and context of hate speech: Rethinking regulation and responses*, pages 417–429.

[7] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma. Deep learning for hate speech detection in tweets. In proceedings of the 26th International Conference on World Wide Web Companion, pages 759{760, 2017.

[8] Suh, J. H. (2016). Comparing writing style feature-basedclassification methods for estimating user reputations in social media. SpringerPLus, 5(1), 1.

[9] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber (2017). Automated hate speech detection and the problem of offensive language. *In Proceedings of the International AAAI Conference on Web and Social Media (ICWSM).*American Association for Artificial Intelligence, Toronto, Canada.

[10] Zeerak Waseem and Dirk Hovy. (2016). Hateful symbols or hateful people? Predictive features forhate speech detection on Twitter. *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*ACL, SanDiego, California, pages 88–93.