2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON) | 978-1-6654-7978-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/NIGERCON54645.2022.9803164

Multi-Label Classification of Hate Speech Severity on Social Media using BERT Model

Bakwa Dunka Dirting
Department of Computer Science
Federal University of Technology, Owerri
Imo State, Nigeria
bakwadunka@gmail.com

Gloria A. Chukwudebe
Department of Electronic Engineering
Federal University of Technology, Owerri,
Imo State, Nigeria
gloria_chukwudebe@ieee.org

Euphemia Chioma Nwokorie
Department of Computer Science
Federal University of Technology, Owerri.
Imo State, Nigeria
cenwokorie@gmail.com

Ikechukwu Ignatius Ayogu
Department of Computer Science
Federal University of Technology, Owerri.
Imo State, Nigeria
ignatius.ayogu@futo.edu.ng

***Abstract*—Detection of offensive and hate speeches on social media using multi-label classification technique is a relatively new fine-grained solution to classification problems. This paper investigates intelligent learning models based on the BERT model for multi-label classification of hate speech. The approach utilized a semi-supervised pseudo-labeling technique to automatically label a newly created multi-social media data which was then augmented and balanced using AugLy and GPT-2 libraries before being used to train the BERT model. Alpha evaluation of the model returned a score of 0.948695 for toxic, 0.946662 for severe toxic, 0.944483 for obscene, 0.946159 for threat, 0.909272 for insult and 0.734659 for identity hate respectively. Examples were ranked and one among such ranked examples gave a probability score of 96%, 89.91% and 80.21% for the top three likely labels. The results compared well with that of the human-annotated severity ranking.**

Keywords— Offensive speech, Hate speech detection, Multi-label classification, Hate speech severity, Deep learning

I. Introduction

Hate speech is hazardous and inimical to healthy communications. To preserve the essence of social networks, the automatic detection of textual conversations that go against the rules of normal communications or are capable of instigating violence calls for continuous research. The Multi-label classification technique is relatively a new fine-grained solution to single classification methods in current solutions.

From Internet live statistics, more than 8,000 tweets are posted per second every day, this eventually amounts to more than 260 billion tweets per year coming from 45% of the world's total population. The staggering amount of social media data coming from various platforms aggregate big data from which information can be extracted through a text mining process. The majority of techniques that were used in the past to solve the problem of hate speech detection from the literature are characterized by a series of limitations. For example, the dominant use of Twitter as a social platform does not generalize well for the term "social media [1]. Another challenge is where text classification and representation suffer linguistic characteristics such as ambiguity, sarcasm, polysemy, negation, and misinterpretation. This affects the generalization benchmark across platforms due to the subjective and complicated nature of the human language as well as the type and styles of communication on social media. Sometimes, what is referred to as hate speech could be seen as freedom of speech by certain persons and vice versa. Therefore, context matters in the overall task of hate speech detection to clear doubts in terms of bias. For this reason, it is important to develop a new system with a sense of intelligence to understand the types of hate [2], [3].

In the past, social networks like Facebook employed 7,500 bystanders to look out for traces of hate on their platforms. Nevertheless, this does not come easy due to the time required to manually filter each post [4], [5].

The traditional problem of single-label classification is concerned with learning from examples, each associated with a single label λ_i from a finite set of disjoint labels (L) defined by L = $\{\lambda_1, \lambda_2, \ldots, \lambda_Q\}$, with Q > 1. For Q = 2, the learning problem is referred to as binary classification problem, if $Q > 2$, the learning problem is referred to as a multi-class classification. The task in multi-label classification is to correctly assign sub-classes to a given class. Thus, in the context of hate speech detection, the multi-label task extends to the problem of ascertaining the level of 'hate-ness' in the hate speech. Theoretically, single-label classification cannot properly scale up to the complexity of the ever-growing social media data as well as correctly identify the correlations that exist between labels. This means the distinctiveness in the subjectivity of hate and the contextual nature of impolite occurrences in social media communications are better handled with multi-label learning. This research seeks to establish the dependences that exist between the hate data (posts and comments) and their probable classes.

II. Literature review

Hate speech detection is a prevalent issue that is increasingly gaining a lot of attention of researchers [6]. For example, social research from the Anti-Defamation League (ADL) division team developed a Pyramid of Hate that explained the context of hate phenomena and their possible consequences. The Pyramid describes toxic behaviors that keep increasing in complexity from the bottom to the top. Behaviors at the next level increasingly become more accepted if the lower levels are tolerated until it reaches a life-threatening situation [7]. The terminologies of hate and offensive speeches are very sensitive. The understanding and subjectivity of hate speech by various annotators can lead to unpredictable performance of the model [8] and [9].

BERT is a pre-trained language representation model that was trained on about 16GB of unlabeled texts data. This includes the entire Wikipedia, Books Corpus with 3.3 billion words and a vocabulary size of 30,522. It is superior to other pre-trained language models such as ULMFiT and ELMo because it was designed to understand the contextual relationship among word sequences bi-directionally using a Masked Language Model (MLM) and next sentence prediction [10].

[11] introduced a hierarchical multi-label classification API, to identify targets, groups, and levels of hate speech for the Indonesian language using Twitter data. Random Forest (RF). Decision Tree (DT), Naïve Bayes (NB), and Support Vector Machines (SVM) were used along with term frequency features such as word n-gram and character n-gram. The research studied five scenarios with different label hierarchies to find the highest accuracy that can be reached by hierarchical classification. They discovered that the SVM algorithm and word unigram feature has the highest accuracy of 68.43%. The total data size used is 31,634, divided into training: 25,307 and testing: 6,327. Their contribution is to the hierarchical form of multi-label classification. However, only Twitter data was used. The results can be improved upon using context-sensitive language models.

The authors utilized the BERT model and recorded good performance with an accuracy of 91.36%. This performance proves to be reliable for automated hate speech detection. An F_1 score measure of 0.81 was recorded against 0.77 and 0.75 for Logistic Regression and Random Forests model,

signifying great improvement in the model's performance. This paper took its motivation from the work of [1], who observed that hate speech detection is a problem across multiple platforms with a lack of competent models for online detection across social multi-platforms. To address this research gap, the authors collected a dataset of 197,566 comments from four platforms: YouTube, Reddit, Wikipedia, and Twitter, with 80% of the comments labeled as non-hateful and the remaining 20% labeled as hateful. This experiment was carried out using several ML classification algorithms, including Logistic Regression, NB, SVM, XGBoost, with feature representations such as bag-of-words, TF-IDF, Word2Vec and Neural Networks and pre-trained language models and ensembles.

In [12], the authors experimented with the transfer learning approach for pre-trained language model, BERT, to detect hate speech on publicly available benchmark datasets. Their result showed that a pre-trained BERT model when fine-tuned on a downstream hate speech task will help to extract its syntactical and contextual features to detect hate speech. However, with the general performance of BERT's transformers, it is unable to de-bias hate speech.

From our analysis, most previous works have used Twitter data for training hate speech models, in contrast, there are many other social media platforms. Also, a single-label classification cannot properly handle the correlation, distinctiveness, and subjectivity inherent in the multi-label classification of hate speech on social media.

III. Methodology

A. Methodology Workflow and Experimental Setup

This research focuses on determining the severity of hate speech on three (3) social media platforms using the BERT models by considering the context of words through high-level features understanding. The approach in this study is shown in Fig. 1.

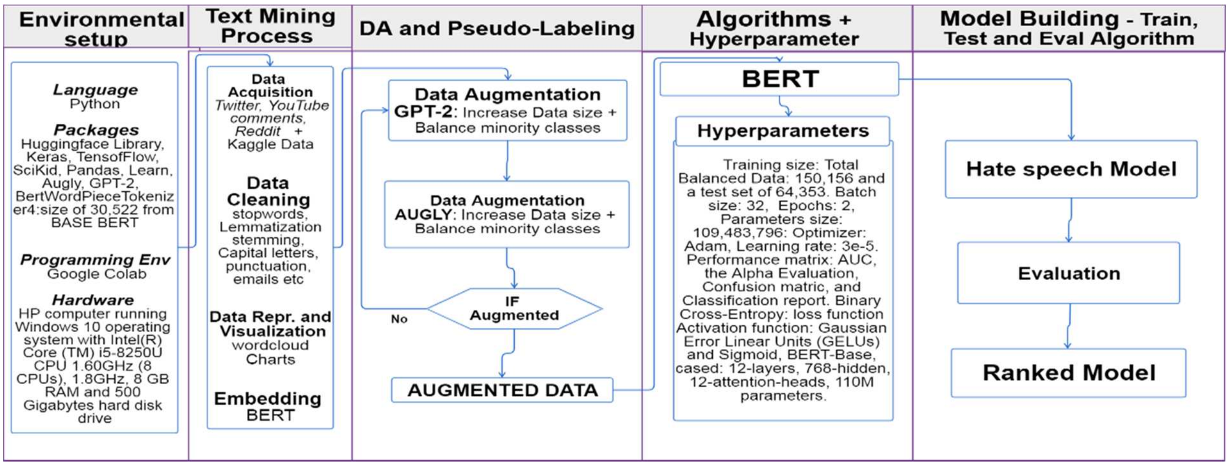

Environmental setup
Text Mining Process
DA and Pseudo-Labeling
Algorithms + Hyperparameter
Model Building - Train, Test and Eval Algorithm
Language
Python
Packages
Huggingface Library, Keras, TensofFlow, SciKid, Pandas, Learn, Augly, GPT-2, BertWordPieceTokenizer4:size of 30,522 from BASE BERT
Programming Env
Google Colab
Hardware
HP computer running Windows 10 operating system with Intel(R) Core (TM) i5-8250U CPU 1.60GHz (8 CPUs), 1.8GHz, 8 GB RAM and 500 Gigabytes hard disk drive
Data Acquisition
Twitter, YouTube comments, Reddit + Kaggle Data
Data Cleaning
stopwords, Lemmatization stemming, Capital letters, punctuation, emails etc
Data Repr. and Visualization
wordcloud Charts
Embedding
BERT
Data Augmentation
GPT-2: Increase Data size + Balance minority classes
Data Augmentation
AUGLY: Increase Data size + Balance minority classes
IF Augmented
No
AUGMENTED DATA
BERT
Hyperparameters
Training size: Total Balanced Data: 150,156 and a test set of 64,353. Batch size: 32, Epochs: 2, Parameters size: 109,483,796: Optimizer: Adam, Learning rate: 3e-5. Performance matrix: AUC, the Alpha Evaluation, Confusion matric, and Classification report. Binary Cross-Entropy: loss function Activation function: Gaussian Error Linear Units (GELUs) and Sigmoid, BERT-Base, cased: 12-layers, 768-hidden, 12-attention-heads, 110M parameters.
Hate speech Model
Evaluation
Ranked Model

Fig. 1. Methodology workflow.

Figure 1, outlines the series of steps that were carried out to achieve the objectives of this paper. The main tasks are experimental setup, text mining process, data augmentation process, pseudo-labeling, algorithm selection, testing, prediction, and evaluation. A reusable process that scrapes and collects data from the multiple social media platforms was realized as suggested by [1]. This is important to address the gap in model generalization across platforms. A single platform provides limited variety and does not generalize effectively in the sense of the term "social media". Secondly, this study aims to diversify the sources of dataset for multilabel classification of hate speech. Majority of existing research drew dataset solely from Twitter.

The Kaggle Toxic speech dataset is one among the two datasets that is used in this study. It contains 159,571 comments. Out of this number, 143,346 representing 89.83% of the comments are clean, fewer comments of size 16,225 representing 10.16% are those that are disproportionately distributed across the remaining 6 classes. For this reason, this study generated additional 30,000 hate/toxic comments from YouTube comments, Reddit, and Twitter. The sum of the data, about 46,225 was augmented to increase the entire set and balance the minority classes (Fig. 2).

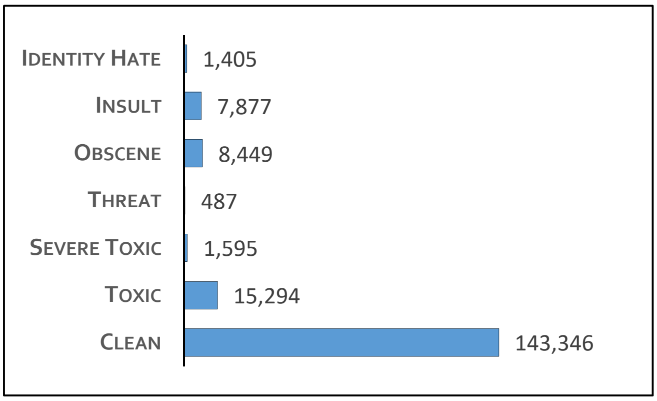


Fig. 2. The distribution of the number of comments per class.

The additional data point generated were conformed to that of the Kaggle dataset using semi-supervised pseudo-labelling. BERT models were then trained to identify the data that belongs to each label and the knowledge was subsequently used to iteratively label the new multi-social media set based on a high probability threshold of 0.7 and above. The labeled data were appended to the entire set and learned the same way until all the labels were learned. This research implemented Data Augmentation (DA) to tackle the problem of data insufficiency and cater for the minority classes, a problem inherent with multi-label classification problems [13], [14]. Similarly, the newly launched AugLy library was used for DA. Both implementation were very useful in adding synthetic test samples. The Final sample of the augmented data was also cleaned and appended to the primary data for use in all the experiments.

The baseline architecture of BERT was used to specifically train the prepared data from scratch. Thereafter, the statistically ranked outcomes were determined from probability scores to determine the hierarchy of severity of hateful instance(s). Label ranking in this study refers to the task of prioritizing the set of predefined labels for severity based on the contextual understanding of BERT for any given set of comments. The ranking model outputs the labels in order of their predicted relevance for automated moderation.

A. Implementation of Data Augmentation (DA) and Pseudo-labelling

GPT-2 and AugLy Facebook library were used in this study to create similar synthetic data to solve the problem of data insufficiency. GPT-2 was designed to produce clean hate comments. Similarly, the AugLy Facebook library was also more useful for the same purpose. However, the pieces of augmented comments are modified by attributes such as overlaying meme-style text, spaces, manipulation of letters and special characters with words.

B. Implementing Pseudo-labelling for Labels Assignment

The fundamental idea of Pseudo-labeling was to train the model with a BERT base-cased classification algorithm separately on the labeled Kaggle data, then use the trained layers or knowledge to predict the labels of the new unlabeled dataset based on a threshold accuracy score. A threshold accuracy score of 0.7 and above was chosen in our experiments based on observations from a random exploratory experiments that were initially conducted. The loss function is given in (1).

$$L = \frac{1}{n}\sum_{m=1}^{n}. \sum_{i=1}^{C}. L\,(y_i^m . f_i^m) + a(t)\cdot\left(\frac{1}{n'}\sum_{m=1}^{n'}. \sum_{i=1}^{C}. L\,(y_i'^{\,m} . f_i'^{m})\right) \qquad (1)$$

Where n is the number of mini-batch for the label data, n' is the unlabeled data. f_i^m is the output units of m, y_i^m is the labels of $f_i'^m$for unlabeled data, $y_i'^{\,m}$ is the pseudo-label for unlabeled data, $a(t)$ is a coefficient for balancing both. The unlabeled loss uses a to control the loss using weights as a function in the unlabeled data during each training epoch. The model initially focuses on the labeled data especially when its performance is not good. The weight gradually increases with the number of epochs over time. α(t), given in (2). The parameters, t, T_1, T_2, α_f, are controllable hyperparameters. They also change based on the data, epoch, and model.

$$\alpha(t) = \begin{cases} 0 & t < T_1 \\ \frac{t-T_1}{T_2-T_1}\,\alpha_f & T_1 \le t < T_2 \\ \alpha_f & T_2 \le t \end{cases} \qquad (2)$$

C. Experimental Setup

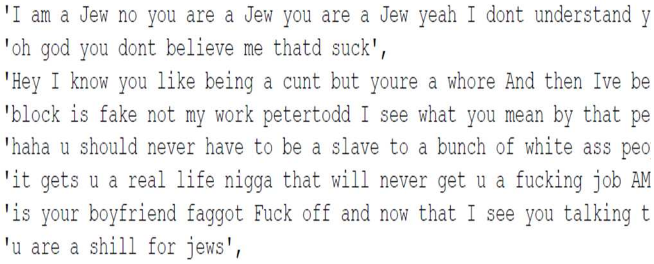
The model used for experiment in this study consists of 12 layers, and 12 attention heads and a total of 100 million parameters.

To conduct the experiments, the model was configured using PyTorch framework for HuggingFace's transformer library for BERT-based implementations. All models were trained using the NVIDIA P100 GPU on Google Colaboratory over a batch size of 32 and a sequence length of 128. Adam was used for optimization, with a learning rate of 3e-5. The binary cross-entropy loss and Gaussian error linear units (GELUs) activation function were used. Classification report, AUC score, and the alpha-Evaluation were the evaluation measures.

IV. Results

A. Result of GPT-2 Library Augmentation

The results of augmentation from GPT-2 is presented. The comments were programmed to be cleaned and conformed to human-like comments. Both results were necessitated to increase the size of the minority classes. Fig. 3. Shows an excerpt of augmented hate comments generated by GPT-2.



```
'I am a Jew no you are a Jew you are a Jew yeah I dont understand y
'oh god you dont believe me thatd suck',
'Hey I know you like being a cunt but youre a whore And then Ive be
'block is fake not my work petertodd I see what you mean by that pe
'haha u should never have to be a slave to a bunch of white ass peo
'it gets u a real life nigga that will never get u a fucking job AM
'is your boyfriend faggot Fuck off and now that I see you talking t
'u are a shill for jews',
```

Fig. 3. Example of Data Augmentation using GPT-2.

B. Classification result

The classification report is shown in Table 1. The results show good evaluation scores.

TABLE 1. THE CLASSIFICATION REPORT.

Category	Precision	Recall	F$_1$-Score	α- Eval
Toxic	0.97	0.9	0.93	0.95
Severe Toxic	0.96	0.96	0.96	0.95
Obscene	0.95	0.95	0.95	0.95
Threat	0.96	0.92	0.94	0.95
Insult	0.93	0.87	0.9	0.91
Identity Hate	0.69	0.19	0.3	0.73

C. α- Evaluation Score

The result obtained using α- Evaluation score is shown in Table. 3. The general results correspond well with most of the hand-annotated examples.

Table 3: *α- Evaluation Results.*

	α- Evaluation Score
Toxic	0.95
Severe Toxic	0.95
Obscene	0.94
Threat	0.95
Insult	0.91
Identity Hate	0.73

D. Hate Speech Severity Results

Fig. 4, shows comment picked randomly from our test set that says:

"Kill Yourself. I mean just look at you, you fucking faggot. You are Indian. You lose. Kill yourself swagfuckingtastic".

The orange bars indicate comments with high probability values in this case representing: *Toxic, Threat, and Insults*. Whereas the comment is less of *Severe toxic, Obscene, and Identity hate* as shown with the blue bars. The ranking hierarchy was automatically decided by the learned representations of the BERT embedding.

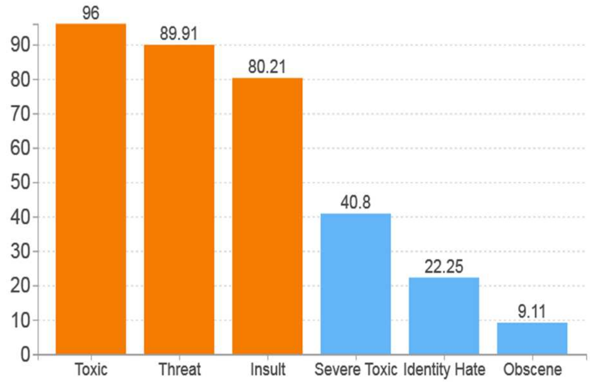


Fig. 4. The result of the severity distribution.

V. Conclusions

In this paper, a method for multi-label classification of hate speech on social media and its severity developed using the BERT model has been described, experimented and evaluated. A new dataset that was created using data scrapped from three social media platforms to increase variability and improve on the representativeness of the dataset. Experimental results show that applying semi-supervised pseudo-labeling and data augmentation were significant for increasing the performance of the BERT model for hate speech severity classification in a multi-label scenario. Further research is necessary to conduct an ablation study that would enable ground truth tests of the label classes and real-time classification of APIs across social media platforms.

References

[1] J. Salminen, M. Hopf, S. A. Chowdhury, S. G. Jung, H. Almerekhi, & B. J. Jansen. "Developing an online hate classifier for multiple social media platforms", Human-centric Computing and Information Sciences, vol. 10, no. 1, 2020, https://doi.org/10.1186/s13673-019-0205-6.

[2] B. Presant, et al., "Hate Speech Classification Using BERT LING 575": "The Cool Kids". 2019.

[3] M. Ioannis, C. Zoe, K. Stamatis, T. Grigorios. "ETHOS: a multi-label hate speech detection dataset. Complex & Intelligent Systems". 2022. 10.1007/s40747-021-00608-2.

[4] H. H. Mohammed, "Multi-label Classification of Text Documents using Deep Learning," unpublished, 2019, p.3

[5] Z. Waseem, & D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter". In Proceedings of the North American chapter of the association for computational linguistics: Human Language Technologies, 2016, Association for Computational Linguistics (ACL), pp. 88–93.

[6] B. Gamback, & U. K. Sikdar, "Using convolutional neural networks to classify hate speech". In Proceedings of the 1st Workshop on Abusive Language Online at ACL, 2017. Vancouver, BC, Canada, 30 July-4 August 2017, pp. 85–90.

[7] PeaceTech Lab, "Social Media and Conflict in Nigeria, A Lexicon Of Hate Speech Terms ADL Education Division: Pyramid of Hate", available at http://www.adl.org/assets/pdf/educationoutreach/Pyramid-of-Hate.pdf.

[8] A. Schmidt and M. Wiegand, “A survey on hate speech detection using natural language processing”. In International Workshop on Natural Language Processing for Social Media, pages 1–10. Association for Computational Linguistics, 2017. doi:10.18653/v1/W17-1101.

[9] N. Gitari, Z. Zuping, H. Damien, and J. Long, “A lexicon-based approach for hate speech detection”. International Journal of Multimedia and Ubiquitous Engineering, 2015 10(10):215–230, doi:10.14257/ijmue.2015.10.4.21.

[10] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". 2018. arXiv:181004805 [cs]. pp. 4171-4186.

[11] F. A. Prabowo, M.O. Ibrohim, & I. Budi, “Hierarchical multilabel classification to identify hate speech and abusive language on Indonesian Twitter”. In 2019 6th International Conference on Information Technology, Computer and Electrical Engineering, ICITACEE, 2019 [8904425]. IEEE Inc.. https://doi.org/10.1109/ICITACEE.2019.8904425.

[12] M. Mozafari, R. Farahbakhsh, & N. Crespi, “A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media” in International Conference on Complex Networks and Their Applications. Springer, 2019, pp. 928–940.

[13] M. Benk, “Data Augmentation in Deep Learning for Hate Speech Detection in Lower Resource Settings”. Unpublished Thesis. Master of Arts der Philosophischen Fakultat der Universitu at Zurich. Retrieved on March 16th from https://www.cl.uzh.ch/dam/jcr:57406b34-02c8-496d-9b95-9968cee3a134/benk_ma_data_augmentation.pdf.

[14] M. S. Ahmed, L. Khan, and N. Oza, “Pseudo-label generation for multi-label text classification”. University of texas at dallas. Conference on Intelligent Data Understanding 2011. P.61.