An approach to detect abusive content incorporating Word2Vec and Multilayer Perceptron

Sayani Ghosal
Computer Science & Engineering
NSUT East Campus (erstwhile A.I.A.C.T.R.), Guru Gobind Singh Indraprastha University
Delhi, India
sayanighosal@gmail.com

Amita Jain
Computer Science & Engineering
Netaji Subhas University of Technology
Delhi, India
amita.jain@nsut.ac.in

Devendra Kumar Tayal
Computer Science & Engineering
Indira Gandhi Delhi University for Women
Delhi, India
dev_tayal2001@yahoo.com

***Abstract*— With the rapid growth of social media text, millions of negative comments are flowing on social webs and social networking sites. Abusive content is harmful to people and societies that can provoke various criminal offenses like hate crimes. Hate speech is also a form of abusive content. An automatic and improved detection system for hate speech can help to reduce this problem. Implicit abusive content requires contextual semantic and syntactical analysis. We propose a novel abusive text detection model with the word2vec model and compositional vector model to analyze text more semantically and syntactically. The proposed model considers the English language dataset for abusive text. The abusive content detection model exhibits achievable performance compare to various deep learning and machine learning classifiers. Among all models, Multilayer Perceptron classifier achieves 86% accuracy compared to other models.**

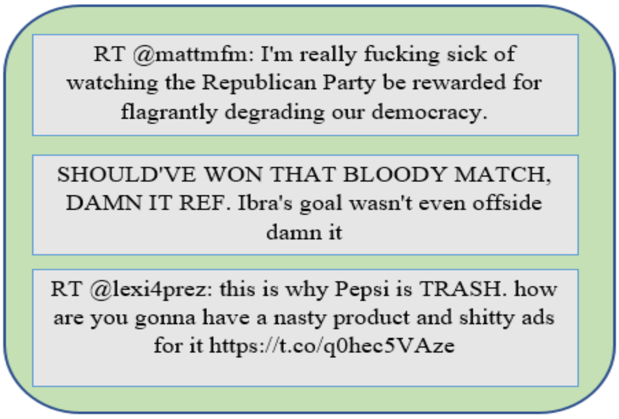
Keywords—Abusive Text Detection, Natural Language Processing, Word2Vec, Compositional Vector Model, Multilayer Perceptron

I. Introduction

Abusive content is an expression that contains slang, dirty and abusive words. Hate speech is also a form of abusive text [1]. In recent times, due to the growth of social webs, social network sites, like Twitter, Facebook, Instagram, and YouTube share malicious, abusive content. Every day, a huge number of abusive posts flow on social media. Abusive language targets individuals or groups of people based on religion, skin color, celebrity, product and politician. Users can target comments on public figures or communities without any fear. So, abusive comments against any person or group should be detected and restricted to maintain a healthy environment.

Fig.1 shows various abusive content extracted from Twitter [2]. First abusive content shows aggressiveness with slang words against one group. The second comment shows aggression against one player and disappointment against a match. The third comment displays rejection against a product and shows abusive emotion against a brand.

Manual detection of such types of abusive comments is impossible from millions of Twitter comments. Conventional deep learning and machine learning techniques are widely applied for abusive text detection [3]. NLP is also an important approach to detect abusive content from social media text [3]. Various lexicon based approaches [4], deep learning models [5], traditional machine learning models [6] and ensemble approaches [7] were already applied to detect abusive content efficiently.



RT @mattmfm: I'm really fucking sick of watching the Republican Party be rewarded for flagrantly degrading our democracy.

SHOULD'VE WON THAT BLOODY MATCH, DAMN IT REF. Ibra's goal wasn't even offside damn it

RT @lexi4prez: this is why Pepsi is TRASH. how are you gonna have a nasty product and shitty ads for it https://t.co/q0hec5VAze

Fig. 1: Sample Abusive content extracted from Twitter[1]

Deep learning models like RNN [8], LSTM [9], CNN [10] and BiLSTM [5] are also popular approaches to detect abusive text. Along with that, some recent models applied hybrid approaches like hierarchical attention-based BiLSTM and Deep CNN approach [11] and LSTM model with CNN and Glove embedding [12]. One recent unsupervised approach with word2vec and cosine similarity detects abusive text from the Twitter dataset [13]. The availability of different types of abusive datasets also encourages researchers for more efficient detection models.

Abusive text detection suffers from various challenges. As per recent research, abusive sentences are grammatically correct and tracing racist comments is difficult [14]. Some recent research identified that the detection of implicit abusive text still suffers [15]. As per recent research, the classification of abusive text is also not able to identify the overlapping contents [16].

Considering the above limitations, we contribute novel abusive content detection research. This is the first abusive content detection research that considers compositional vector model and word embedding for tweet vector computation. The contributions are as follows – firstly, analyzing abusive content with the features of word2vec and compositional vector model. This research considers abusive text from Twitter. Secondly, this research considers MLP classifiers for efficient abusive text detection. It also compares with various deep learning and machine learning models.

[1] https://twitter.com

2022 IEEE Bombay Section Signature Conference (IBSSC) | 978-1-6654-9291-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/IBSSC56953.2022.10037274

978-1-6654-9291-1/22/$31.00 ©2022 IEEE

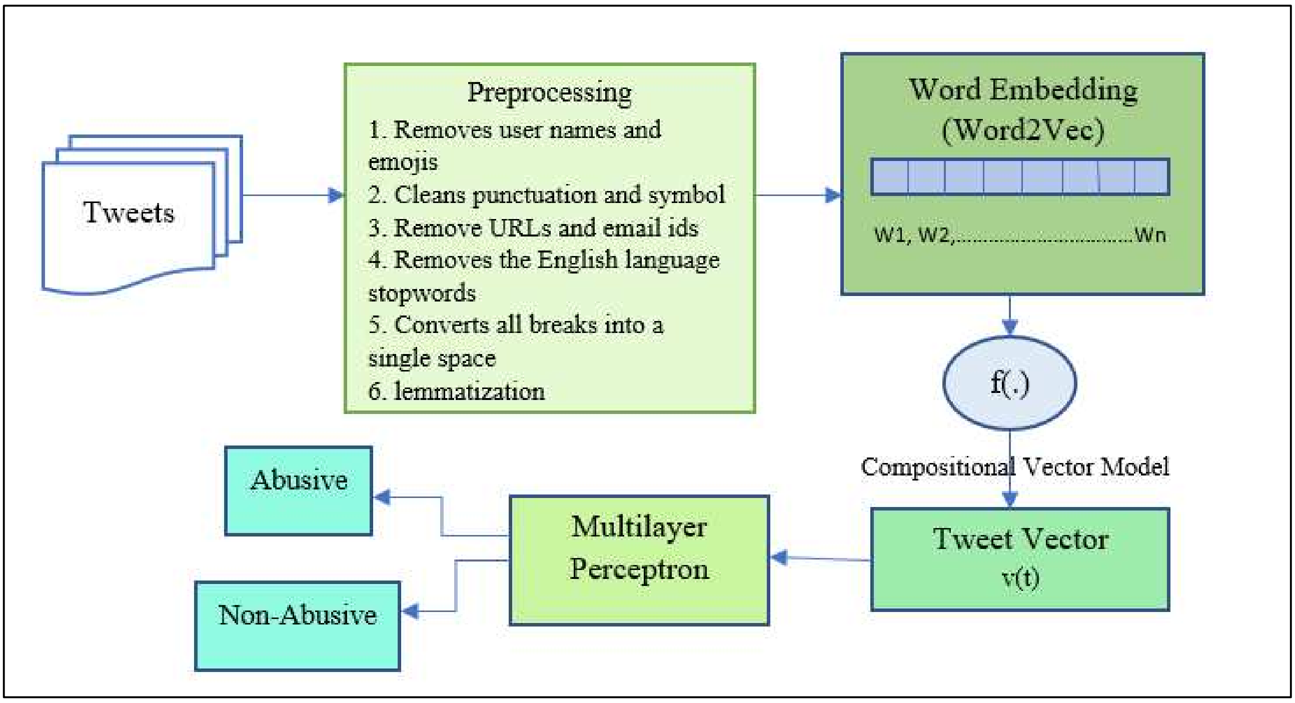

Tweets
Preprocessing
1. Removes user names and emojis
2. Cleans punctuation and symbol
3. Remove URLs and email ids
4. Removes the English language stopwords
5. Converts all breaks into a single space
6. lemmatization
Word Embedding (Word2Vec)
W1, W2,...............................Wn
f(.)
Compositional Vector Model
Tweet Vector v(t)
Multilayer Perceptron
Abusive
Non-Abusive

Fig. 2: Flow diagram of abusive content detection model

This research is organized as follows, section 2 represents the methodology for proposed abusive content detection with all features and the classification model. The result section displays the experimentation of the proposed method and all findings in section 3. Finally, this research concludes in section 4.

II. Methodology

This proposed abusive content detection approach considers the word embedding model and compositional vector model to extract the tweet vector from each tweet. This research extracts abusive text by using tweet vector features and deep learning classifier multilayer perceptron (MLP). Fig.3 exhibits the proposed architecture of the abusive text detection model. The abusive text detection approach consists of four steps – preprocessing, word embedding, compositional vector model, and multilayer perceptron.

A. *Preprocessing*

To improve the accuracy of the abusive text detection model, this research considers various preprocessing steps. Initially, it removes user names and emojis from tweets. It also cleans punctuation and symbol from each tweet. Preprocessing steps remove URLs and email ids. Along with the above steps, it also removes the English language stopwords applying NLTK library [17]. It converts all breaks into a single space for all tweets. Finally, preprocessing steps consider lemmatization to reduce inflected forms of words.

B. *Word Embedding*

The This abusive text detection approach considers the word2vec model to represent word vectors for each tweet. The Word2Vec model [18] can efficiently extract the semantic and syntactical relationship between words. Two types of word2vec models exist for large text word embedding – CBOW and skip-gram. It was observed in various textual applications that CBOW performs well for large text. CBOW model considers words from tweets and embeds them in the same semantic space through vectors.

Each tweet is denoted as tϵT. Each tweet contains n number of words $\{W_{1,}W_2 \ldots. W_n\}$. This CBOW model sets 200 vector dimensions for abusive text detection.

C. *Compositional Vector Model*

Various text research considers the average of word vectors to form tweet vectors or sentence vectors. This abusive content detection research computes sentence vectors or tweet vectors from word vectors using the Bi- function. Bi-function of the compositional vector model [19] represents each tweet vector from word vectors. This compositional function captures bi-gram information from tweets. It is also able to extract word interaction of tweets by hyperbolic tangent function. This function is continuous in nature and produces output in the range of [-1, +1].

Let t implies a tweet with n word vectors $\{\{W_{1,}W_2 \ldots. W_n\}\}$ with m dimensions. So, bi-function defined as:

$$v(t) = \sum_{i=1}^{n} f([W_{i-1,} + W_{i,}]) \qquad (1)$$

v(t) is the combined vector representation of each tweet. Element-wise vector addition is represented as f([p+q]]) where two vectors are p and q. Function f(.) is defined as hyperbolic tangent function

$$f(x) = tanh(x) = \frac{e^x - e^x}{e^x + e^x} \qquad (2)$$

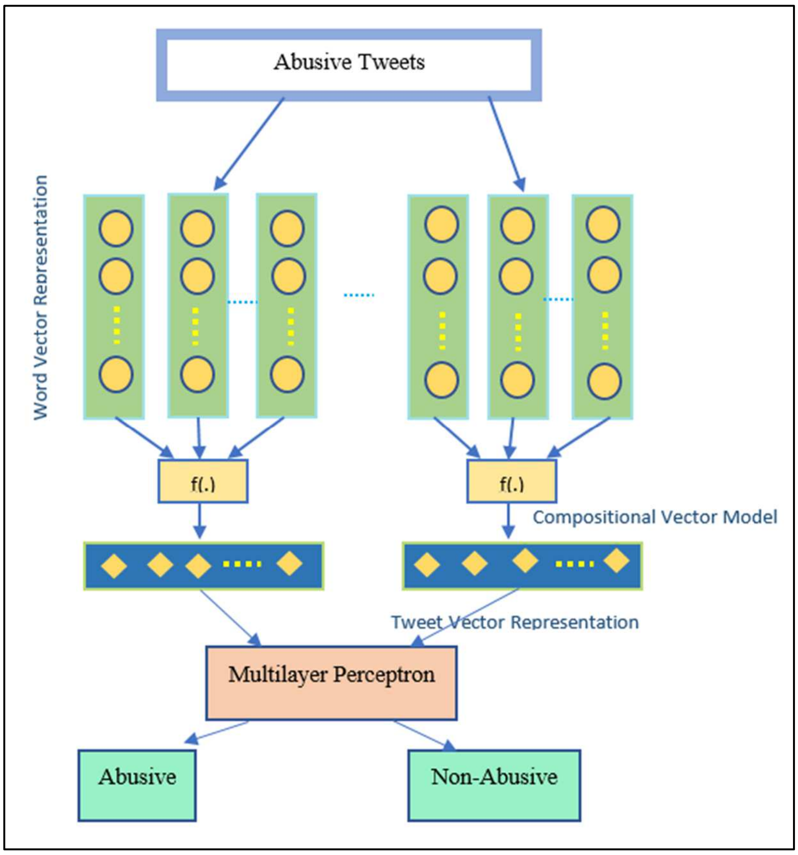

Abusive Tweets
Word Vector Representation
f(.)
f(.)
Compositional Vector Model
Tweet Vector Representation
Multilayer Perceptron
Abusive
Non-Abusive

Fig. 3: Proposed model architecture

D. *Multilayer Perceptron*

The MLP classifier [20] is a feed-forward artificial neural network that consists of input, output and weight layers. MLP classifier considers activation function for computation of each layer and backpropagation algorithm is used to adjust weights of MLP classifier. This classifier also provides nonlinear mappings of input vectors and output vectors. Proposed abusive text detection research considers this MLP classifier for tweet vectors v(t) receives from the compositional vector model layer. The result section exhibits that the MLP classifier outperforms compare to other classifiers for abusive content detection.

III. Result and Analysis

Dataset: The abusive content dataset extracted from Twitter[2] social networking sites [21]. This dataset contains 60903 tweets. This research considers only the English language dataset. Dataset label with abusive and non-abusive text. We consider 70% dataset for training and 30% dataset for testing.

Evaluation metrics: This proposed abusive content detection research considers precision (P), recall (R), weighted F1 score (F1) and accuracy (Acc). All these evaluation metrics consider for analyzing the performance of proposed abusive text detection research.

A. *Results and analysis*

This proposed research compared with various word embedding features and different deep learning and machine learning classifiers.

TABLE I. Experiment results comparison with various features

Sl. No.	Features	P	R	F1	Acc
1	Bag of words	0.68	0.69	0.69	0.68
2	Tf-idf	0.72	0.72	0.72	0.72
3	Paragraph Embedding	0.79	0.79	0.79	0.78
4	fastText Embedding	0.80	0.80	0.80	0.80
5	Word2Vec	0.84	0.84	0.83	0.84
6	Word2Vec + Bi- function	0.86	0.86	0.84	0.86

In Table I of abusive content detection research, we consider various word embedding models to show the performance of our proposed model. All the embedding model shows good accuracy whereas the word2vec model shows better compare to other features. It also observed that the word2vec model with bi-function (compositional vector model) outperforms compare to all features.

TABLE II. Compare with both word2vec model

Sl. No.	Features	P	R	F1	Acc
1	CBOW + Bi-function	0.86	0.86	0.84	0.86
2	Skip-gram + Bi- function	0.84	0.84	0.84	0.84

[2] https://twitter.com/?lang=en-in

Table II shows the comparison of two word2vec models – CBOW and skip-gram with bi-function. It was observed that both the models achieve same F1 score but the accuracy of the model CBOW is better compared to the skip-gram vector model with bi-function.

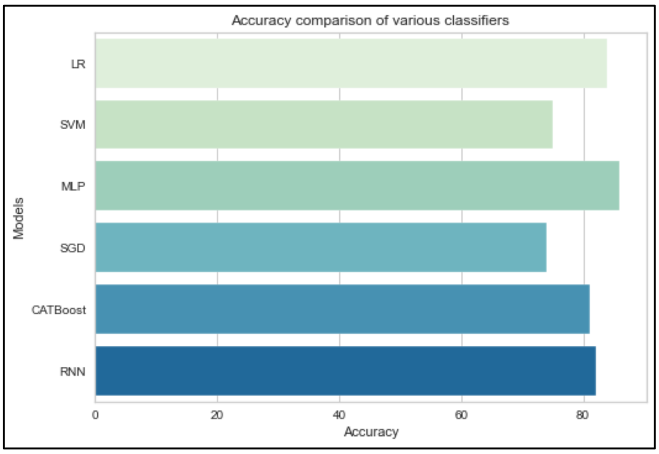


Fig. 4: Accuracy comparison with various models

Fig 4 represents comparisons of various classifiers with the word2vec model and bi-function feature sets. It was observed that the proposed model outperforms compare to the other machine learning and deep learning classifiers. All these classifiers consider with features set word2vec model and bi-function. Here logistic regression (LR) also shows 0.84 Acc. SGD and RNN both deep learning models also not able to achieve good Acc and SVM achieves less Acc compare to all other classifiers.

TABLE III. EXPERIMENT RESULTS COMPARISON WITH VARIOUS CLASSIFIERS

Sl. No.	Features	P	R	F1
1	LR (Logistic Regression)	0.83	0.83	0.83
2	SVM	0.74	0.73	0.74
3	SGD	0.73	0.73	0.73
4	CATBoost	0.80	0.80	0.81
5	RNN	0.82	0.83	0.82
6	MLP (Multilayer Perceptron)	0.86	0.86	0.84

Table III exhibits the comparison with various classifiers with the word2vec model and bi-function feature sets. This table considers P, R and weighted F1 parameters for comparison. Here logistic regression (LR) and MLP classifier shows high weighted F1 score compare to other classifiers. SGD and RNN models also able to achieve good F1 score and SVM and SGD achieves less F1 score compare to all other classifiers. It was observed that due to large tweets deep learning classifiers perform well.

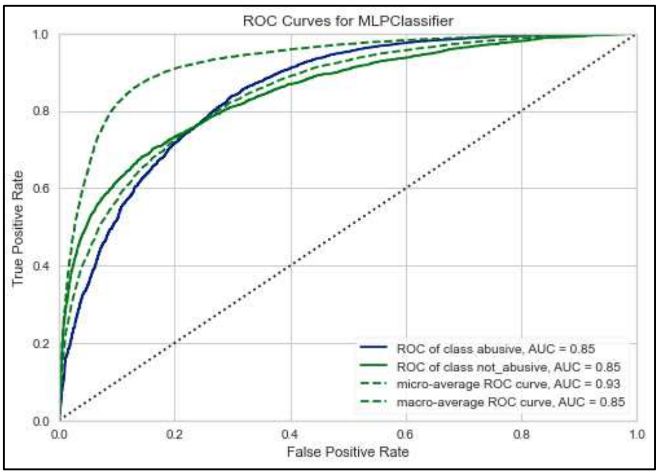


Fig. 5: ROC curve for Abusive text detection

Fig 5 presents the ROC curve of binary classification of abusive content detection. The ROC curve exhibits the comparison of true positive rates and false positive rates. Here 1 is considered abusive and 0 is considered non-abusive.

IV. CONCLUSION AND FUTURE SCOPE

The proposed research efficiently detects abusive content from a large Twitter dataset. It was observed that the Word2vec embedding model enhanced the classification of implicit abusive content by contextual analysis. The compositional vector model also efficiently identified the bi-gram information from tweets. The combination of both features detected abusive content syntactically and semantically. We consider a large Twitter English language dataset for this abusive content detection research. It was observed that the MLP classifier improves accuracy compared to all other deep learning and machine learning classifiers and achieves 86% accuracy and 0.84 weighted F1-score for abusive content detection. In the future, the proposed model may enhance with emotion and sentiment features. This research also may improve with various domain-specific lexicons. This proposed research may also enhance by applying sarcasm and humor identification from various low-resource languages.

REFERENCES

[1] S. Ghosal, & A. Jain. "Research Journey of Hate Content Detection From Cyberspace". In Natural Language Processing for Global and Local Business. IGI Global. pp. 200-225, 2021

[2] Twitter, https://twitter.com, Accessed October 2022

[3] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. AbdelMajeed and T. Zia. "Abusive language detection from social media comments using conventional machine learning and deep learning approaches". Multimedia Systems, pp.1-16, 2021.

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

[5] G. I. Sigurbergsson, & L. Derczynski. "Offensive language and hate speech detection for Danish". arXiv preprint arXiv:1908.04531. 2019

[6] S. C. Eshan, & M. S. Hasan. "An application of machine learning to detect abusive bengali text". In 2017 20th International conference of computer and information technology (ICCIT). IEEE. pp. 1-6, December 2017.

[7] R. Pelle, C. Alcântara, & V. P. Moreira. “A classifier ensemble for offensive text detection”. In Proceedings of the 24th Brazilian Symposium on Multimedia and the Web. pp. 237-243, October 2018.

[8] J. H. Park, & P. Fung. “One-step and two-step classification for abusive language detection on twitter”. arXiv preprint arXiv:1706.01206, 2017

[9] P. Badjatiya, S. Gupta, M. Gupta, & V. Varma. “Deep learning for hate speech detection in tweets”. In Proceedings of the 26th international conference on World Wide Web companion. pp. 759-760, April 2017.

[10] Y. Lee, S. Yoon, & K. Jung. “Comparative studies of detecting abusive language on twitter.” arXiv preprint arXiv:1808.10245. 2018

[11] S. Khan, M. Fazil, V. K. Sejwal, M. A. Alshara, R. M. Alotaibi, A. Kamal, & A. R. Baig. “BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection”. Journal of King Saud University-Computer and Information Sciences, 34(7), pp. 4335-4344, 2022.

[12] M. Anand, & R. Eswari. “Classification of abusive comments in social media using deep learning”. In 2019 3rd international conference on computing methodologies and communication (ICCMC), IEEE. pp. 974-977, March 2019.

[13] H. S. Lee, H. R. Lee, J. U. Park, & Y. S. Han. “An abusive text detection system based on enhanced abusive and non-abusive word lists”. Decision Support Systems, 113, pp. 22-31, 2018.

[14] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, & Y. Chang. “Abusive language detection in online user content”. In Proceedings of the 25th international conference on world wide web. pp. 145-153, April 2016.

[15] S. Ghosal, & A.Jain. “Analysis of Misogynistic and Aggressive Text in Social Media with Multilayer Perceptron.” In Emerging Technologies in Data Mining and Information Security. Springer, Singapore. pp. 589-596, 2023

[16] S. Sadiq, A. Mehmood, S. Ullah, M. Ahmad, G. S. Choi, & B. W. On. “Aggression detection through deep neural model on twitter.” Future Generation Computer Systems, 114, pp. 120-129, 2021.

[17] E. Loper, & S. Bird. “ Nltk: The natural language toolkit.” arXiv preprint cs/0205028, 2002.

[18] T. Mikolov, K. Chen, G. Corrado, & J. Dean. “Efficient estimation of word representations in vector space.” arXiv preprint arXiv:1301.3781, 2013.

[19] K. M. Hermann, & P. Blunsom. “Multilingual models for compositional distributed semantics.” arXiv preprint arXiv:1404.4641, 2014.

[20] E. B. Baum. “On the capabilities of multilayer perceptrons.” Journal of complexity, 4(3), pp. 193-215, 1988.

[21] Abusive Text Detection: From Traditional Machine Learning to Deep Learning Approaches, https://github.com/Shawon-Lodh/Abusive-Text-Detection, Accessed September, 2022.