IJERST

International Journal of Engineering Research and Science & Technology

ISSN : 2319-5991

www.ijerst.com

Email: editor@ijerst.com or editor.ijerst@gmail.com

ISSN 2319-5991 www.ijerst.com
Vol. 17, Issue 2, 2024

Hate speech classification on social media using a service framework

[1]DR.B.GOHIN, [2]ATKURI VINEETHA

[1](Associate Professor), MCA, Swarnandhra College

[2]MCA, scholar, Swarnandhra College

ABSTRACT

It is indeed a challenge for the existing machine learning approaches to segregate the hateful content from the one that is merely offensive. One prevalent reason for low accuracy of hate detection with the current methodologies is that these techniques treat hate classification as a multiclass problem. In this article, we present the hate identification on the social media as a multilabel problem. To this end, we propose a CNN-based service framework called "HateClassify" for labeling the social media contents as the hate speech, offensive, or nonoffensive. Results demonstrate that the multiclass classification accuracy for the CNN-based approaches particularly sequential CNN (SCNN) is competitive and even higher than certain state-of-the-art classifiers. Moreover, in the multilabel classification problem, sufficiently high performance is exhibited by the SCNN among other CNN-based techniques. The results have shown that using multilabel classification instead of multiclass classification, hate speech detection is increased up to 20%.

1.INTRODUCTION

The rise of social media has made it easy for people to express themselves emotionally. On the other hand, the broad use of social media under the guise of free speech, has resulted in the proliferation of hate speech. Between 2014 and 2016, the amount of hate speech posted on social media skyrocketed by more than 900%, according to a recent USA Today article. According to a survey (https://www.pewresearch.org/internet/2014/10/22/onlineharassment/), 73% of internet users have witnessed online harassment, and

ISSN 2319-5991 www.ijerst.com
Vol. 17, Issue 2, 2024

40% have experienced it themselves. According to the Council of Europe's Protocol to the Convention on Cybercrime, language used to "spread, incite, promote, or justify racial hatred, xenophobia, anti semitism or other forms of hatred based on intolerance, including intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants, and people of immigrant origin" is considered "hate speech." However, under the First Amendment's free speech clauses, hate speech is protected in the United States. Concerning the definition of "hate speech," on their individual social media websites, companies like Google, Facebook, and Twitter all have their own standards. The various social media platforms have different opinions on how to handle hate speech and other forms of objectionable content. Just one of the three major social media platforms—Twitter, Facebook, and Google—does not have any restrictions on hate speech. When it comes to hate speech and outright threats, Twitter makes a distinction. As far as Twitter is concerned, "one-sided" accounts whose main goal is to harm others are the only ones whose hostile conduct is taken into consideration. Twitter continues to take heat for having rules that are too open-ended, even though the company insists that no one is exempt. Facebook, Twitter, YouTube (owned by Google), as of May 31, 2016, Microsoft and the European Union have signed a voluntary code of conduct to eliminate hate speech. The topic of hate was brought up after the CEO of Facebook was questioned about the company's policy regarding the recognition and reporting of hate speech and material.speech on social media came to a lot of attention lately. The CEO of the business made it clear in his comments that Facebook's present method of identifying hate speech fails to accurately recognize the range of emotions expressed. This is due to the fact that hate speech content is subjectively defined by various people. Offensive and hate speech was identified as an issue in a number of earlier studies, such as Del Vigna et al.1. Hate speech and offensive speech were, however, defined differently by Davidson et al. 2. According to the study's authors, individuals routinely use very harsh language. A multiclass classification issue including hate, offensive, and nonoffensive speech was therefore proposed as a solution to the challenge of hate speech categorization. The classification of talks offered by Davidso et al.2 is one with

ISSN 2319-5991 www.ijerst.com
Vol. 17, Issue 2, 2024

which we concur. However, rather than seeing the hate speech issue as a multiclass problem, we view it as a multilabel one. Even human specialists have struggled to discern between hate speech and offensive speech due to the subtle differences between the two. Thus, resolving disputes between disputing parties will never be as simple as firmly identifying one class. By framing the issue as a multi-label problem, we were able to improve the accuracy of hate speech and offensive speech detection. Hate Classify is a proposed service architecture that uses a mix of machine learning and crowdsourcing to identify hate speech and objectionable content on social media sites. This article mostly adds the following.
As a service to social media firms, we provide a framework for hate speech and objectionable content identification.

Instead than having individual organizations police hate speech regulations on social media platforms, the suggested framework uses a crowd-sourced method to identify hate speech.

By treating the problem of detecting hate speech as a multi-label classification problem, we are able to achieve a level of classification accuracy that is high enough. How can hate speech on social media be detected?the Hate Classify framework's multi label categorization improves the process by 20%. What follows is an outline of the remaining content of this piece. In the "Related Work" part, we talk about the linked projects. The "Framework for Hate Speech Detection" part presents the service framework. The study is wrapped up in the "Conclusions" section, which presents the findings of multiclass and multi-label classification in addition to comparisons with the most recent technology approaches.

2.LITERATURE SURVEY

"Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Learning Model" Published by Zhang et al. in 2020For the purpose of detecting This study presents a deep learning model that combines Gated Recurrent Units (GRUs) and Convolutional Neural Networks (CNNs) to combat hate speech on Twitter. Tweets' local and sequential information are both captured by the model., allowing it to attain state-of-the-art performance.

The second piece is "A Survey on Hate Speech Detection using Natural Language Processing" by Fortuna et al. (2018).

In order to identify hate speech, this review compiles a number of different methodologies that make use of Natural Language Processing (NLP). It discusses the benefits and drawbacks of several methodologies, including lexicon-based approaches, deep learning techniques, and machine learning.

"Deep Learning-Based Hate Speech Detection on Social Media Texts" was published in 2016 by Nobata et al.

In order to identify hate speech on social media, the writers provide a deep learning method. With the help of CNNs and LSTM networks, they are able to autonomously learn features from text input and attain competitive performance.

"A Survey on Automated Hate Speech Detection in the Social Media" (Burnap and Williams, 2015).

Methods such as lexicon-based approaches, machine learning algorithms, and hybrid approaches are included in this review, which offers an overview of automated hate speech detection strategies in social media. It delves into the difficulties of hate speech identification and suggests ways forward for study.

"Hate Speech Detection: Is It Finally Over?" "The Challenging Case of Long Tail on Twitter" Importantly, it shows how important it is to have reliable models that can identify hate speech in many languages and settings. "An Ensemble Method for Hate Speech Detection in Twitter" (Mehdad et al., 2016).

Using a combination of different classifiers and characteristics, this research suggests an ensemble approach to identify hate speech on Twitter. For better identification accuracy, the method integrates lexicon-based characteristics, word embeddings, and language patterns.
7) ""The Problem of Offensive Language and Automated Hate Speech Detection"

In 2017, Davidson et al.

The writers go into the difficulties of automated detection of hate speech with an emphasis on the distinction between hate speech and insulting language. to differentiate between insulting language and hate speech, they suggest a multi-class categorization method.

3. EXISTING SYSTEM

The work on the hate speech detection mostly revolves around finding the best features that

ISSN 2319-5991 www.ijerst.com
Vol. 17, Issue 2, 2024

can be used in text classification algorithms. The basic features that are used by most of the authors in their studies are n-grams and Bag-of-Words (BoW). Warner et al.3 argued that hatred against different groups can be categorized with the usage of small set of high frequency words. Chen et al.4 used n-grams with syntactic rules, such as user's writing style. Hosseinmardi et al.5 used n-grams along with the number of comments for the images. Length of a tweet, geographical location, and gender information of the tweeting person were used along with the n-grams for hate speech detection by Waseem and Hovy.6 Finding the grammatical usage of hate content has also gained popularity among the researchers. Van Hee et al.7 used the sentiment features along with the n-grams and the BoW for studying and detecting hate speech. Xu et al.8 used n-grams with the Part-Of-Speech tagging (POS tagging) to study bullying traces on the social media. Davidson et al.2 used TF-IDF weighted unigram, bigrams, trigrams, sentiment score of the tweet.

number of hashtags, retweets, URLs, characters, words, and syllables in each tweet as the feature set. To overcome the problem of sparsity due to short length of texts in tweets or online comments during hate detection, numerous researchers have utilized the concept of word generalization. Warner and Hirschberg3 used Brown Clustering technique for word generalization. Unlike Brown Clustering that assigns word to exactly one cluster, latent Dirichlet allocation (LDA) predict the probabilities of word in different clusters.

Xiang et al.9 used the LDA for word generalization. Recently, several distributed word representations, termed as the word embedding have been developed for word generalizations. The word embedding takes the large text as the input and develops a vector space of words. The word vectors are placed in such a manner that words with similar context are placed closer to each other. Zhong et al.10 used word2vec (a word embedding technique) along with the BoW and hate effectiveness score to detect the hate speech. Paragraph2vec another word embedding technique was studied for hate speech detection against the BoW approach by Djuric et al. For classification, state vector machine12,3-5,7-9 and logistic regression (LR)2;6;9 have outperformed the other techniques for the hate speech detection studies. Nobata et al.13 preferred Vowpal

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

Wabbit's regression model over other models. Mehdad and Tetreault14 have used recurrent neural network (RNN) models for hate speech detection.

Disadvantages

- In the existing work, the system did not implement Multilabel Classification Results.
- This system is less performance due to lack of CNN model which is for hate classification is sequential convolutional neural network model (SCNN).

3.1 PROPOSED SYSTEM

› The system presents a framework for detection of hate and offensive speech as a service for social media companies.

› Contrary to the social media platforms where the policies regarding hate speech are regulated by the specific organizations, the proposed framework employs a crowd-sourced approach for hate speech identification.

› The problem of hate speech detection is presented as multi label classification problem and sufficiently high classification accuracy is achieved.

› The multi label classification used in Hate Classify framework yields 20% improvement in detection of hate speech on social media.

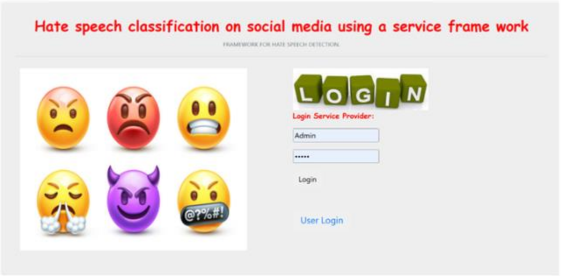
Advantages

(i) an offline training module- The offline training is a periodic job that takes the tweets and labels the tweets tagged by different people.

(ii) online hate and offensive speech detection module.

4. OUTPUT SCREENS

Home Page

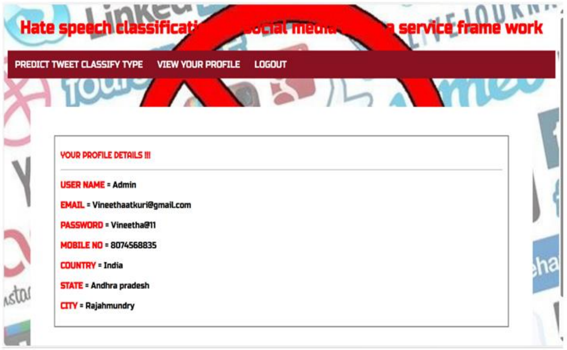

Hate speech classification on social media using a service frame work
FRAMEWORK FOR HATE SPEECH DETECTION
LOGIN
@?%#!
Login
User Login

Login Service Provider:

Admin

•••••

View profile page


Hate speech classificati social media a service frame work
PREDICT TWEET CLASSIFY TYPE
VIEW YOUR PROFILE
LOGOUT

YOUR PROFILE DETAILS !!!

USER NAME = Admin

EMAIL = Vineethaatkuri@gmail.com

PASSWORD = Vineetha@11

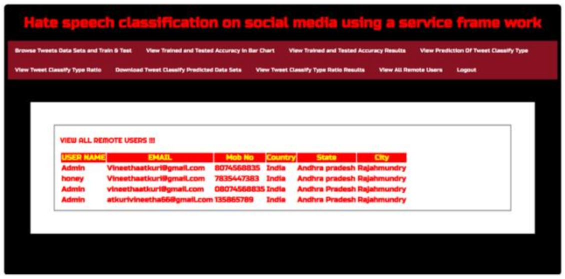
MOBILE NO = 8074568835

COUNTRY = India

STATE = Andhra pradesh

CITY = Rajahmundry

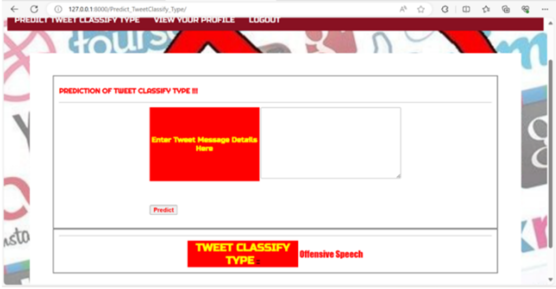
View Remote Users


Hate speech classification on social media using a service frame work
Browse Tweets Data Sets and Train & Test
View Trained and Tested Accuracy in Bar Chart
View Trained and Tested Accuracy Results
View Prediction Of Tweet Classify Type
View Tweet Classify Type Ratio
Download Tweet Classify Predicted Data Sets
View Tweet Classify Type Ratio Results
View All Remote Users
Logout

VIEW ALL REMOTE USERS !!!

USER NAME	EMAIL	Mob No	Country	State	City
Admin	Vineethaatkuri@gmail.com	8074568835	India	Andhra pradesh	Rajahmundry
honey	Vineethaatkuri@gmail.com	7835447383	India	Andhra pradesh	Rajahmundry
Admin	vineethaatkuri@gmail.com	08074568835	India	Andhra Pradesh	Rajahmundry
Admin	atkurivineetha66@gmail.com	135865789	India	Andhra Pradesh	Rajahmundry

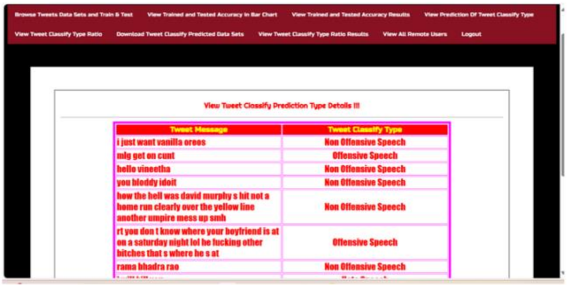
Output


127.0.0.1:8000/Predict_TweetClassify_Type/
PREDICT TWEET CLASSIFY TYPE
VIEW YOUR PROFILE
LOGOUT
Enter Tweet Message Details Here
Predict

PREDICTION OF TWEET CLASSIFY TYPE !!!

TWEET CLASSIFY TYPE : Offensive Speech

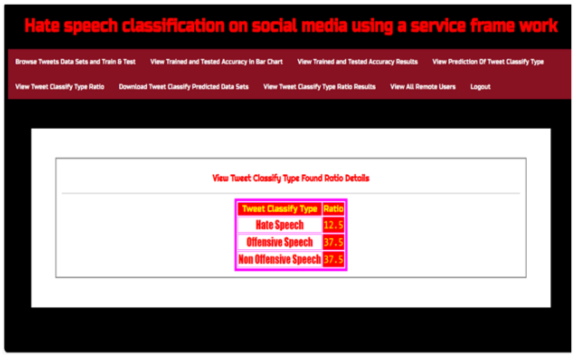
Predict type classification


Browse Tweets Data Sets and Train & Test
View Trained and Tested Accuracy in Bar Chart
View Trained and Tested Accuracy Results
View Prediction Of Tweet Classify Type
View Tweet Classify Type Ratio
Download Tweet Classify Predicted Data Sets
View Tweet Classify Type Ratio Results
View All Remote Users
Logout

View Tweet Classify Prediction Type Details !!!

Tweet Message	Tweet Classify Type
i just want vanilla oreos	Non Offensive Speech
mlg get on cunt	Offensive Speech
hello vineetha	Non Offensive Speech
you bloddy idoit	Non Offensive Speech
how the hell was david murphy s hit not a home run clearly over the yellow line another umpire mess up smh	Non Offensive Speech
rt you don t know where your boyfriend is at on a saturday night lol he fucking other bitches that s where he s at	Offensive Speech
rama bhadra rao	Non Offensive Speech

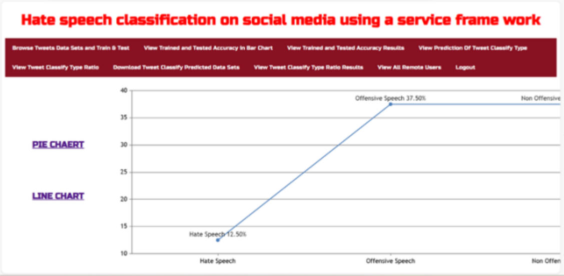
Tweet ratio


Hate speech classification on social media using a service frame work
Browse Tweets Data Sets and Train & Test
View Trained and Tested Accuracy in Bar Chart
View Trained and Tested Accuracy Results
View Prediction Of Tweet Classify Type
View Tweet Classify Type Ratio
Download Tweet Classify Predicted Data Sets
View Tweet Classify Type Ratio Results
View All Remote Users
Logout

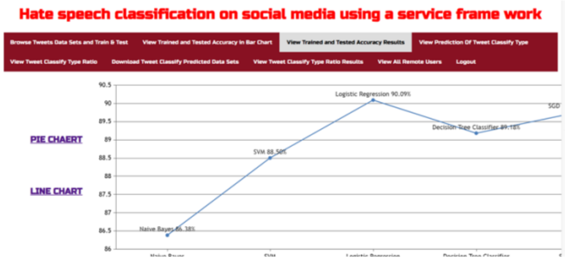
View Tweet Classify Type Found Ratio Details

Tweet Classify Type	Ratio
Hate Speech	12.5
Offensive Speech	37.5
Non Offensive Speech	37.5

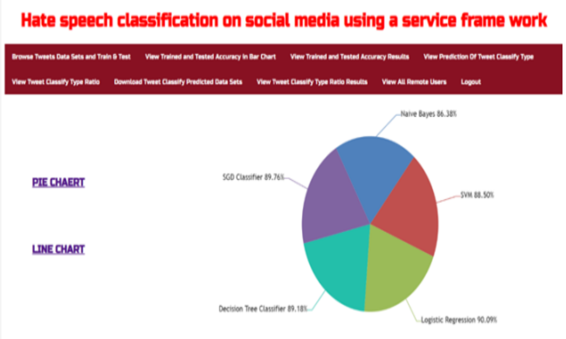
Line chart



pie-chart



Bar-chart



ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue 2, 2024

5. CONCLUSION

To identify hate speech on social media, we introduced a service architecture named Hate Classify in this paper. When it comes to unacceptable textual speech or material, the Hate Classify methodology uses a crowdsourced approach that lets social media users vote. We used convolutional neural networks (CNNs) to assess classification performance, and our experiments show that CNN models, and the SCNN in particular, outperform various state-of-the-art methods when it comes to classification accuracy. By framing the hate speech classification issue as the multi label classification problem, this study makes a significant addition to the field. It is feasible to use convolutional neural network (CNN) methods for social media hate speech categorization, according to the experimental findings obtained using these methods for multiclass and multilabel classification**.**

6.REFERENCES

1. F. Del Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi, "Hate me, hate me not: Hate speech detection on Facebook," in Proc. 1st Italian Conf. Cybersecurity, 2017, pp. 86–95.

2. T. Davidson, D. Warmsley, M. Macy, and I.Weber,"Automated hate speech detection and the problem of offensive language," in Proc. 11th Int. AAAI
Conf. Web Social Media, 2017, pp. 512–515.

3. W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in Proc. 2nd Workshop Lang.
Social Media, 2012, pp. 19–26.

4. Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," in Proc. IEEE Int. Conf. Privacy, Secur., Risk Trust Int. Conf. Soc. Comput., 2012,pp. 71–80.

5. H. Hosseinmardi, S. A.Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Analyzing labeled cyberbullying incidents on the instagram social network," Social Inform.,T. Y. Liu, C. N. Scollon, andW. Zhu, Eds., 2015, pp. 49–66.