# Abstract

As the increasing number of social media user emerging day by day from various backgrounds and different diverse moral codes to today’s wildly popular platforms,a space for hate space has emerged. Although social media platforms favours communication and information sharing,these are also used to launch harmful campaigns against individuals or specific groups.We aim at containing and preventing such hate campaigns.With the increasing amount of hate speech online,methods that automatically detects hate speech is very much required.Therefore,our goal is to use Natural Language processing to detect hate speech.We propose a Recurrent Neural Network structurer that will serve as a feature extractors which will be explicitly effective for capturing the context and the semantics of hate speech.We will evaluate our methods on the largest collection of hate speech dataset from twitter.Our classifier will assign each tweet as one of following categories:hate,offensive and neutral.More specifically,It will distinguish hate speech form normal text and can achieve higher classification quality than current state-of-art algorithms.

# Introduction

As the technology is growing day by day the communication through internet is also increasing. Every remote parts of the world are connected now to each other through massive number of social sites. Every minute, there are 510,000 comments generated on Facebook [2]. Also 350,000 tweets generated on Twitter in every minute [3]. Usage of social media is increasing with a very high rate and it is enlightening us in many ways. On the other hand, due to the easy access to the social media in the name of freedom of speech some people are spreading hatred. Moreover, people are belonging different community; ethnicity, religion and financial status are all in the same social sites. As a result, if someone has intensions to hurt any group he/she can use the social platforms to do the evil works. According to American Constitution “Hate speech is speech that attacks a person or a group on the basis of attributes such as race, religion, ethnic origin, national origin, sex, disability, sexual orientation, or gender identity. It is a matter of concern that people are involving themselves into hatred so much now a days by sharing controversial topics against different group of people. Therefore a detection process for hatred content is highly needed for all social sites. Our target is to make a hatred detection system which wills automatically checks contents before uploading for the social sites and classify into three classes (Hate, offensive, neutral). The task is really challenging because of the complexity of natural language, also people use different ways or many different words to represent same meaning. For this reason we will use deep learning methods to detect hate speech. To implement our system we will use dataset from kaggle. We will modify our dataset using nltk model. For our system we will implement feature extractor using Word2vec model.

# Literature Review

From [4], we have found that they mainly used convolutional neural network (CNN), bag of words and word2vec for featu. They preferred CNN over recurrent neural networks. But we prefer RNN over CNN because CNN is very time consuming. Moreover, it needs a huge amount of data to train any model whereas RNN needs less than CNN. In addition, CNN works on the current inputs only but RNN works on both current and previously received inputs. RNN has a memory to memorize previous input. One more thing is that CNN cannot handle sequential data while RNN can. In sequential data, sample size is never fixed and data is evaluated when they are collected. In [5], they used glove instead of word2vec. Though both works almost same but there is still a small difference between word2vec and glove. Word2vec is a predictive model whereas glove is a count based model. But we prefer word2vec for our model as it is more reliable word embedding technique than glove. It learns words from a large corpus more quickly than glove. From [6], we got to know some difficulties that every researchers face. They noted from their reviewed papers that many researchers failed to publish their weight initialization methods. For model’s training, they initially used keras library with tensor flow backend but later they switched to theano to solve the issue with reproducibility of weight initializations.

# Work Flow

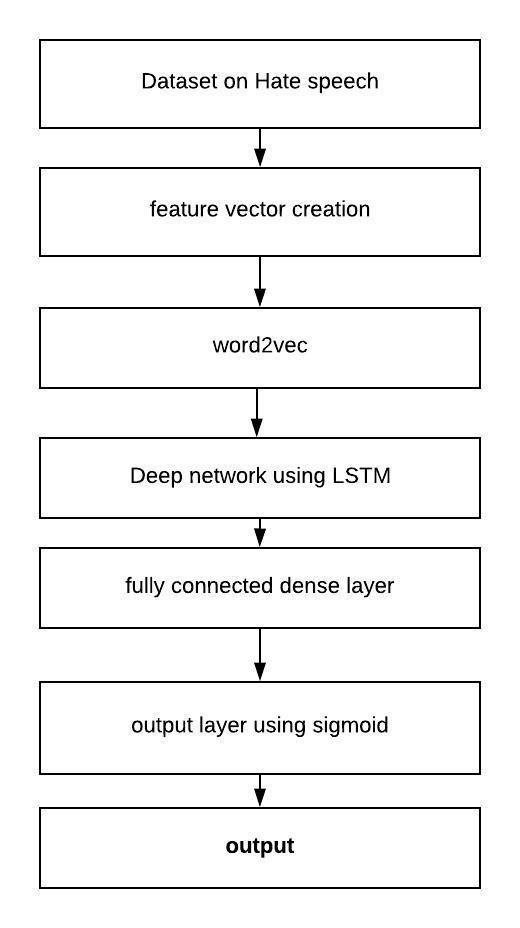


Fig: workflow

# Expected Result

# Conclusion

Automated detection of hate language has become a key challenge in recent day.The anonymity and mobility of social media platform has increased the propagation of hate speech in recent years. In this paper the main aim is to detect hate speech using natural language technique. We have developed a supervised machine learning classifier for hateful and antagonistic content on social media platforms. Firstly, we tried to classify and understand the inherent and contextual meaning of hatred speech from the analyzed dataset through feature extraction.Then we propose a RNN based method for capturing the implicit features that are essential for classification.Finally,our methods will be thoroughly evaluated on a large dataset of twitter to show that they can be particularly effective on detecting and classifying hateful content.

# References

[1]  Nockleby, John T. (2000), “Hate Speech” in *Encyclopedia of the American Constitution*, ed. Leonard W. Levy and Kenneth L. Karst, vol. 3. (2nd ed.), Detroit: Macmillan Reference US, pp. 1277–79. Cited in "Library 2.0 and the Problem of Hate Speech," by Margaret Brown-Sica and Jeffrey Beall, [*Electronic Journal of Academic and Special Librarianship*, vol. 9 no. 2 (Summer 2008)](http://southernlibrarianship.icaap.org/content/v09n02/brown-sica_m01.html#_edn2).

[2] Zephoria.com, 2018. [Online]. Available: https://zephoria.com/top-15- valuable-facebook-statistics/. [Accessed: 22- Jun- 2018].

[3] “Twitter Usage Statistics - Internet Live Stats”, Internetlivestats.com, 2018. [Online]. Available: http://www.internetlivestats.com/twitterstatistics/. [Accessed: 22- Jun- 2018]

[4] <https://beta.vu.nl/nl/Images/werkstuk-biere_tcm235-893877.pdf>

[5]<https://www.researchgate.net/publication/317300025_Deep_Learning_for_Hate_Speech_Detection_in_Tweets>

[6] <http://www.lrec-conf.org/proceedings/lrec2018/pdf/292.pdf>