Textual Content Moderation using Supervised Machine Learning Approach

Revati Ganorkar[1], Gaurang Suki[2], Shubham Deshpande[3], Mayur Giri[4], Araddhana Deshmukh[5]
[1,2,3,4] Student, Department of Computer Science, Smt. Kashibai Navale College of Engineering., Pune
5Aarhus University, Herning, Denmark
revatisganorkar@gmail.com, gaurang210498@gmail.com, shubhamdeshpande123@gmail.com, mgiri6612@gmail.com, , aaradhna.deshmukh@gmail.com[5]

ABSTRACT
By the increasing use of Social Networking Sites, a huge amount of data is generated on daily basis. This data contains a plethora of hate speech and offensive content which makes a negative impact on society. Various tech giants such as Facebook[1] and Microsoft have been using manual content moderation techniques on their website. But even this has a negative effect on content moderators reviewing content across the world. In order to tackle this issue, we have proposed an efficient automated textual content moderation technique which uses supervised machine learning approach.

KEYWORDS
Social Networking Sites, content moderation, hate speech, offensive words, text classification

1. INTRODUCTION

Social Networking Sites have gained a considerable amount of popularity in recent years. It has totally changed people's way of communication and sharing of information.

People use different means for communication (Example: text messages, images, audio clips, video clips, etc) This information shared on social networking sites may contain some data which might be offensive to some people. Also, the shared media may contain some illegal information which can spread the wrong message in the society.

In [4], authors have observed that the increase in the use of social media and Web 2.0 are daily drawing more people to participate and express their point of views about a variety of subjects. However, there are a huge number of comments which are offensive and sometimes non-politically correct and so must be hindered from coming up online. This is pushing the services providers to be more careful with the contents they publish to avoid judicial claims. This work proposes the use of automatic textual classification techniques to identify and only allow to go online harmless textual posts and other content.

Different sites use different methods to moderated the textual content. SNS like Facebook[1], Twitter[2] manually moderate the content whereas Linkedin[3] automatically removes the content after reported by a certain number of users. But, manual moderation of content requires manpower and the moderators have to go through a lot of mental stress while moderating the data. Some of the cases where moderators suffered from extreme stress are discussed here. In [5] content moderators alleged Facebook[1] that it failed to keep its moderators safe as they developed post-traumatic stress and psychological trauma from viewing graphic images and videos.

In another incident [6], two employees at Microsoft filed a lawsuit against Microsoft as they were forced to view content that

inhumane which led to severe post-traumatic stress disorder. Thus, manual moderation of abusive content is malicious for the person moderating the content as it causes harmful effects on them.

Therefore there is a need for an efficient technique to monitor hate speeches and offensive words on social networking sites.

2. LITERATURE SURVEY

In [7], the paper includes moderation of multimodal subtleties such as images or text. The authors develop a deep learning classifier that jointly models textual and visual characteristics of pro-eating disorder content that violates community guidelines. For analysis, they used a million photos, posts from Tumblr. The classifier discovers deviant content efficiently while also maintaining high recall (85%). They also discuss how automation might impact community moderation and the ethical and social obligations of this area.

In [8], the proposed system is designed for open source operating system windows or Linux. The implementation of this system is based on PHP framework. MySQL database is used for storing the datasets by configuring the LAMP server in Ubuntu and WAMP server in windows. Also the configuration of PHPMyAdmin. Ubuntu helps to perform various tasks such as creating, modifying or deleting databases with the use of a web browser. Dream viewer is being used for the system development. For recommendation generation, latest version of Apache is used. To configure Wamp with windows environment the integration of Wamp server in windows is done. To make the Web environment scalable it is being integrated with PHP and Wamp. Initially, for the testing purpose, a Phase one development is being established on localhost.

In [9], various techniques applied regarding with data processing, such as weighting of terms and the dimensionality reduction. All these techniques were studied in order to model algorithms to be able to mimic well the human decisions regarding the comments. The results indicate the ability to mimic experts decision on 96.78% in the data set used. The classifiers used for comparison of the results were the K-Nearest Neighbors and the Covalent Bond Classification. For dimensionality reduction, techniques for the extraction of terms were also used to best characterize the categories within the data set.

As SNSs have become of paramount relevance nowadays, many people refuse to participate in or join them because of how easy it is to publish and spread content that might be considered offensive. In [4], the approach accurately identifies inappropriate content based on accusers' reputations. Analysis of reporting systems to assess content as harmless or offensive in SNSs.

3. GAP ANALYSIS

Not all the data generated from SNS can be considered as normal. It contains a considerate amount of data that can be considered as offensive and hateful. Manual content moderation is effective but requires a considerate amount of manpower and sometimes it can be traumatic for humans to examine such inappropriate content. Hence, in recent days some organizations have come up with effective techniques which can be

used for filtering inappropriate content. Following table summarizes all the different techniques used by different organizations to rectify this illegal data.

Reporting Systems	**Automatic vs human intervention**
Udd	Hate reports are automatically filtered
Work, Blue and hoffman	Content withdrawn depending on owner's reputation
Facebook	Manual review of content on social media
Linkedin	Automated withdrawal after reported by fixed no. of user.
Twitter	Manual review of content on social media and also uses automated data.

Table 1: Content Moderation Techniques

Most of the organizations manually monitor the content. Because of this people are exposed to offensive content which sometimes can be hostile for the person monitoring the data and can cause mental stress. There is a need for a system that will automatically monitor offensive content and reduce the manual workload. Thus, we are proposing a system which will automatically monitor SNS for malign content with the help of machine learning.

4. PROPOSED SYSTEM

Automatic content moderation can be achieved with the help of traditional natural language processing techniques coupled with supervised classification learning. Using the association between these two methods, the model for offensive and hateful text detection is proposed. The proposed model is designed to achieve more efficiency in illegal text classification performance.

The main aim of the proposed model is to eliminate the need for manual content moderation. This can be effectively achieved by utilizing techniques of natural language processing and machine learning that when trained with appropriate data, predicts a nearly accurate outcome.

The proposed model is composed of the following core components as shown in Figure 1.

1. **Natural Language Processing**:- It is responsible of taking textual data as input and apply series of natural language processing techniques so that it can be processed by text classifier. Here, sentences are filtered and converted into a vector of numbers.
2. **Training**:- Twitter corpus is given to Natural Language Processing component which converts it into a set of vectors. These vectors and pre-assigned labels are used for construction and training of the classifier model. The model obtained is then improved with parameter tuning. The parameter tuning method used here is 10-fold cross-validation.
3. **Classifier model**:- During training, classifier model is constructed from the vectorized sentences prepared by Natural Language Processing component and label (Offensive/Normal) which are already present in the dataset. Further, this trained classifier model is used for predicting a given sentence whether it's offensive or not. Classifier predicts the outcome

accurately and precisely. For this purpose, 2 algorithms are compared for their classification performance.

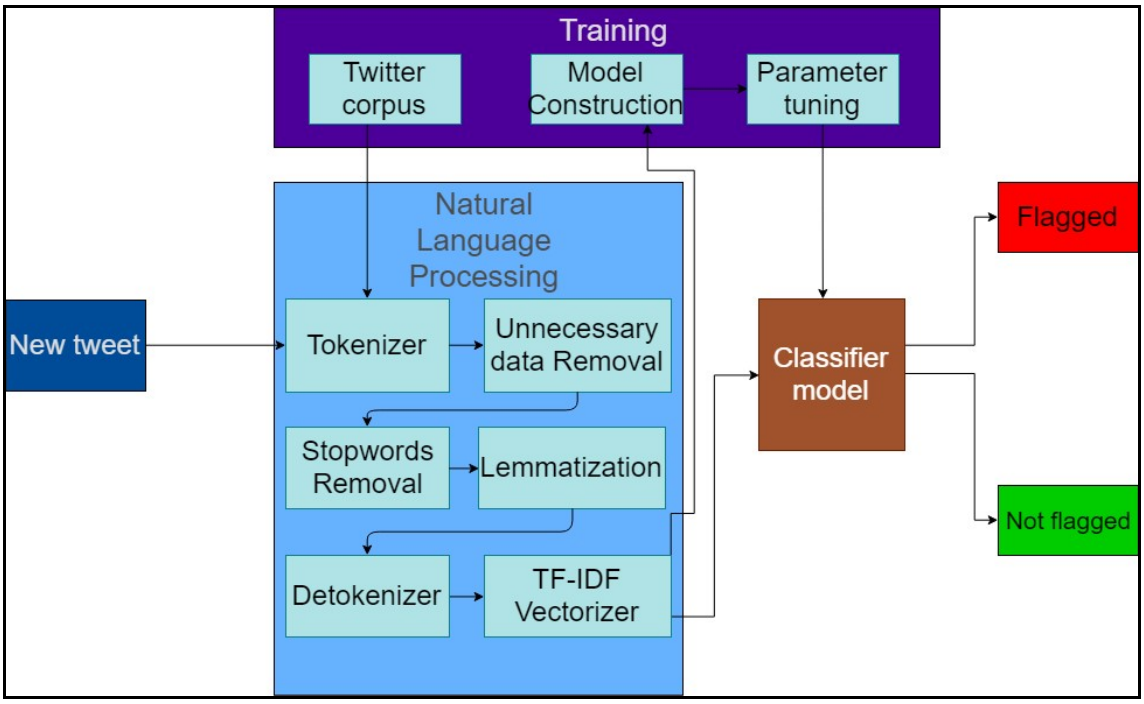

Figure 1: Proposed Architecture

Tweets contain unnecessary data such as stop words, emojis, usernames. This kind of data does not contribute much in the classification and hence, we need to filter out this data as well as normalize it into a suitable format so that it can be used for training the classifier for classifying the unknown text data. An Individual tweet is taken and is then tokenized into words. These tokens are then used to determine unnecessary data such as emoji and usernames. Furthermore, unnecessary symbols and stopwords are removed in order to reduce the data volume.

The main task is to normalize the data. Hence the aim is to infer the grammar independent representation of a given tweet. Lemmatization is used to find out the the lemma of each token. After this, all the filtered tokens for one tweet are collected together for further processing.

The vectorization algorithm used in the proposed model is TF-IDF vectorization. The reason to choose this particular vectorization technique is that the dataset used for the experimentation a contains large number of tweets containing offensive words which dominate the small number of regular tweets. As TF-IDF assigns the score depending upon the occurrence of a term in a document, this seems to be the best choice.

The classifier model is then trained on a collection of pairs containing vectorized tweets and whether they are offensive or not. Supervised classification is used in this proposed system is able to then learn from these tweets and can classify a new tweet.

After training, a new tweet is given to the model, it will repeat all the above steps

except training the model. After going through these steps, vectorized representation of a sentence is obtained. This vectorized representation is then given to previously trained classifier model as input and it classifies the tweet depending on its content.

5. MATHEMATICAL MODEL

The Proposed model can be represented in mathematical model as follows -

Term frequency inverse document frequency (**TF-IDF**) of words in given corpus is calculated by

$$\boldsymbol{tfidf(t,a,D) = tf(t,a)\ x\ idf(t,D)}$$

...(1)

Where,

t - terms

a - individual document

D - collection of document

tf - term frequency i.e. number of times words appear in each document

idf - inverse document frequency calculated by

$$idf(t,D) = log\ \frac{|D|}{1 + \{d \in D :\ t \in a\}}$$

Using (1) all equation are vectorized.

Let V_i represent vectorized sentence i, then general classifier is represented using

$$\hat{y}_i = C(V_i)$$

Where

$\hat{y}_i$: predicted outcome

C : classifier function

Here, we used 2 classifier models. (Bernoulli Naive Bayes and Bagged SVM) for performance comparison

1.) Naive Bayes -

argmax(

$$(P(V_i|y) = P(i|y)x_i + (1 - P(i|y))(1 - x_i)$$

)

2.) Bagged Support Vector Machines -

As given in [12], Support Vector Machines can be bagged as

$$H(d_i) = sign(\sum_{m=1}^{M} \alpha_m H_m(d_i))$$

where,

H_m : Sequence of classifiers

m : 1,....,M

M : Number of classifiers in bagging

α: Learning parameter

6. RESULT & DISCUSSION

We used dataset developed by [10] and further modified it to fit the needs for

classification of the proposed system. This dataset originally contained 3 categories:

1)Normal tweets

2)Offensive tweets

3)Tweets containing hate speech

Only 2 categories are used for the experimentation:- Normal tweets and effensive tweets. Hate speech which also contained offensive tweets are filtered and are treated as offensive tweet only.

The proposed model is implemented in Scikit-learn library[11] in order to obtain results. Following table shows the comparison of various predictive metrics for 2 models which are used for the training.

Results	**Bernoulli Naive Bayes'**	**Bagged SVM**
Accuracy	0.9292543021	0.9492245592
Precision	0.9439205955	0.9700460829
Recall	0.9726412682	0.968805932
F1-Score	0.9580657348	0.9694256108

Table 2: Performance metrics for Bernoulli Naive Bayes' and Bagged SVM

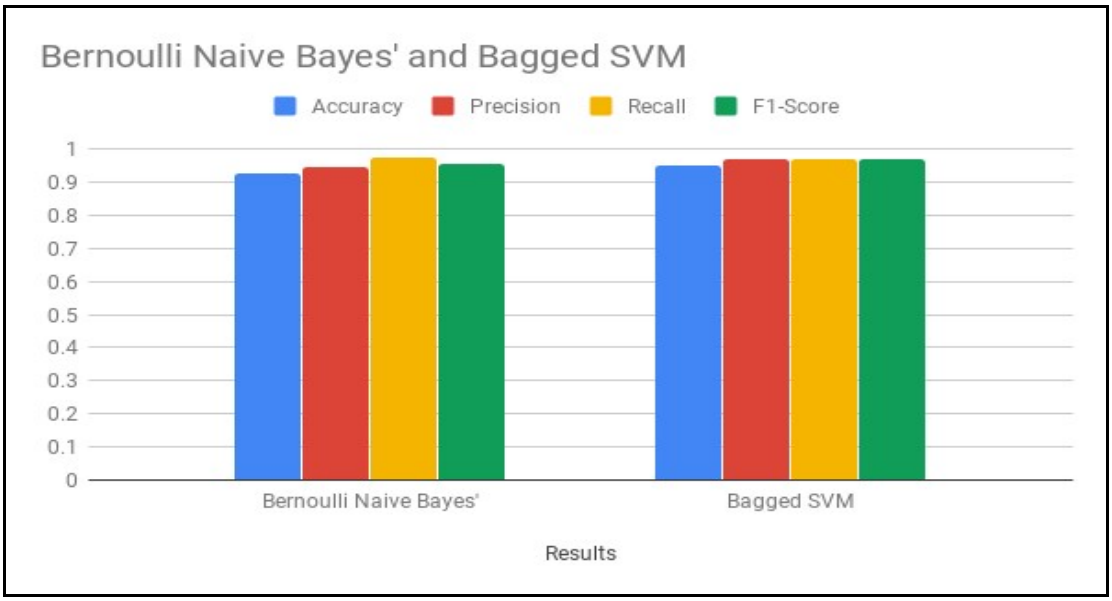

Figure 2 : Bar chart for different metric comparison between the two models

From Figure 2, it can be inferred that both models yield almost same accuracy but by considering other metrics, Bagged SVM performs better than Bernoulli Naive Bayes'.

7. FUTURE WORK

Traditionally content moderation is done manually. This manual work can be reduced using the proposed system. Currently, the proposed system is for textual data but in the future, this can be extended to images, videos, and audio. Further, an efficient model with higher efficiency can be used to classify text data more effectively. Additionally, the algorithm to find out what is wrong with the content can also be designed. Manual

Moderators will be less exposed to hate speeches and offensive if such models are implemented on large scale.

8. CONCLUSION

This system mainly focuses on categorizing text data in two categories namely offensive and normal. This will help content moderators to review less offensive data. Content moderation process will be automated by the use of a machine learning technique.

REFERENCE

[1] Facebook-https://www.facebook.com/ [Access Date: 19 Dec 2018].

[2] Twitter-https://twitter.com/ [Access Date: 19 Dec 2018].

[3] LinkedIn-https://in.linkedin.com/ [Access Date: 19 Dec 2018].

[4] Marcos Rodrigues Saúde, Marcelo de Medeiros Soares, Henrique Gomes Basoni, Patrick Marques Ciarelli, Elias Oliveira. "A Strategy for Automatic Moderation of a Large Data Set of Users Comments". In 2014 XL Latin American Computing Conference (CLEI) (2014,September).

[5] Facebook's 7,500 Moderators Protect You From the Internet's Most Horrifying Content. But Who's Protecting Them. https://www.inc.com/christine-lagorio/facebook-content-moderator-lawsuit.html [Access Date: 19 Dec 2018].

[6] Moderators who had to view child abuse content sue Microsoft, claiming PTSD.

https://www.theguardian.com/technology/2017/jan/11/microsoft-employees-child-abuse-lawsuit-ptsd [Access Date: 19 Dec 2018].

[7] Stevie Chancellor, Yannis Kalantidis, Jessica A. Pater, Munmun De Choudhury, David A. Shamma. "Multimodal Classification of Moderated Online Pro-Eating Disorder Content". In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (Pg. 3213-3226) on ACM (2017,May).

[8] Sanafarin Mulla, Avinash Palave, "Moderation Technique For Sexually Explicit Content". In 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT) at International Institute of Information Technology (I2IT), Pune (2016,September).

[9] Félix Gómez Mármol,Manuel Gil Pérez ,Gregorio Martínez Pérez. "Reporting Offensive Content in Social Networks: Toward a Reputation-Based Assessment Approach". In IEEE Internet Computing Volume 18 , Issue 2 , Mar.-Apr. 2014.

[10] Davidson, Thomas and Warmsley, Dana and Macy, Michael and Weber, Ingmar. "Automated Hate Speech Detection and the Problem of Offensive Language". In proceedings of the 11th International AAAI Conference on Web and Social Media 2017, (Pg. 512-515).

[11] Scikit-learn: A module for machine learning.

https://scikit-learn.org [Access Date: 19 Dec 018].

[12] Kristína Machová, František Barčák, Peter Bednár, "A Bagging Method Using Decision Trees in the Role of Base Classifiers" in Acta Polytechnica Hungarica, Vol.3, No.2, 2006, 121-132, ISSN 1785-8860.