Vulgar And Spam Comment Identification Using Gluon Natural Language Processing And Convolution Neural Networks

By

Rufyda Jahan
13201077

A thesis submitted to the Department of Computer Science and Engineering in partial fulfillment of the requirements for the degree of
Bachelor of Computer Science and Engineering

Department of Computer Science and Engineering
BRAC University
April, 2019

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. I/We have acknowledged all main sources of help.

Student's Full Name & Signature:

Rufydajahan
13201077

Approval

The thesis/project titled “Vulgar and Spam Comment Identification Using Gluon Natural Language Processing and Convolutional Neural Networks” submitted by

1. Student-Rufyda Jahan (13201077)]

of Spring, 2019 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. Computer Science and Engineeringon 25/04/2019.

Examining Committee:

Supervisor:
(Member)

_____

Dr. Iftekharul Mobin
Assistant Professor, CSE
BRAC University

Departmental Head:
(Chair)

_____

Professor Dr. Md. Abdul Motalib
Chairperson, CSE
BRAC University

Abstract/Executive Summary

With the advancement of technology, the virtual platform and social media have become an important part of people's daily life. The social media allows users to communicate, to express their feelings and to discuss on various topics. But cyber bullying and vulgar or toxic comments become an alarming problem in social media. In this paper, a system has been proposed to detect and classify vulgar comments using convolutional neural networking with gluon natural language processing. The system can classify toxic, hate speech, insult, obscene, threat and normal text by training from a huge dataset obtained from Kaggle, extracting comments from YouTube, facebook through extension of browser. The system shows 95.4% accuracy to detect and classify the vulgar comments.

Keywords: Comment Classification, CNN, Natural Language Processing.

Acknowledgement

Initially we would like to thank the almighty for enabling us to initiate our research, to put our efforts and conclude it.

I offer genuine and heartiest appreciation to my regarded Supervisor Dr. Iftekharul Mobinfor his contribution and support in leading the research and preparation of the report. His involvement, inclusion and supervision have inspired me and acted as a huge incentive all through our research.

I'm grateful to the resources, seniors, companions which have been indirectly but actively helpful with the research. We would also like to acknowledge the help we got from different assets over the internet; particularly from fellow researchers' work.

Table of Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Motivation

Internet negativity has perpetually been a hot topic. The obscurity and therefore the sense of distance of people's web presence have inspired to expose themselves freely. As a result, the popularity of internet specially social medium like facebook, twitter, what's app, youtube have risen too high nowadays. But the negative portion of social media has become a new trend of abuse, cybercrime, offensiveness. Vulgarity and Spam can be any bothersome and obliged conduct that could possibly break the security approaches of any arrange security. Nowadays, any vulgar message or fake text sent by spammers to draw the authentic customers. It may be diverse objectives of spammers for sending these spam and disgusting messages like for publicizing any item,cyber tormenting, annoying, threating. There are different sorts of vulgarism and spam as depicted beneath [1]:

Toxic email: Email Spam is a kind of spam sent through email. The connections present in sent sends may delude clients to locales having phishing or malevolent properties. The spammers play in all respects cleverly in the event of email spamming. They collect different email addresses from visit rooms, diverse, sites and so on. furthermore, pitch it to different spammers.

Vulgar comments: Comment Spam is a sort of post in which spammer posts harsh and hostile information in organizing destinations. In comment'sspam , spammer dissipates the spam as remarks on online journals , discussions , Wikipedia and so forth. Instant messenger spam: Instant Messenger Spam gives its clients a majority of catalog of its everything clients having scientific information like name , age , sexual orientation and so on. Spammers searching for

such vulnerable information, assembles all subtleties and by marking in, sends unconstrained and spontaneous messages which may contain any infections , noxious connections and so forth.

Junk fax : Junk fax is similar to email spam. The main distinction lies is that spam in junk fax is obtained as faxes through fax transmission. These are fundamentally utilized for swelling in commercials.

Spontaneous Text Messages : Spontaneous Text Messages is a type of Mobile Spam. This spam type focuses on the message box of any client on his cell phone in any case, this sort is less common than email spam.

Social Media Spam : Social Networking Spam is a sort in which spammed content is posted on social systems. Spam can be as remarks, tweets, talks, pictures and so forth. Fundamentally, above clarified types can be considered inside Social Networking Spam. Now, governments of different countries are trying to prevent cybercrime, bullying and most importantly vulgarity and spam from the internet. Our proposed system has designed in a way to classify the vulgar and spam comments from social media comment so that internet can be again user friendly to all sort of people.

1.2 Literature Review

C. S. Srivastava et al [2] presented capsule network based aggressive comment classifier. The proposed worked was based on focal loss along with single model capsule network which achieves 98.46% accuracy on the on their first dataset and mentioned to apply on TRAC dataset which is Hindi-English combined dataset of aggressive and non-aggressive comment with RNN classifier where α and γ was 2 and 0.25 respectively.

C. Rădulescu et al [3] et al took an integrated approach to identify spams by combining Machine Learning, Natural Language Processing and URL examination methodologies. The combined methodology has shown better and more exact results than when each modularity was applied individually. M. McCord et al detected vulgar comments on Twitter in [4] applying the api method from tweeter. The authors used Random Forest Classifier algorithm to detect vulgar comments or response where they got 95.7 F1 score and 95.7%.

For identifying spam remarks applying Natural Language Processing Techniques Kandasamy et al [5] integrated three approaches- the implementation of URL analysis, supervised machine learning techniques, natural language processing to classify the vulgar and toxic comment from social media where they got 94% accuracy.

Carreras et al built up a strategy to demonstrate that the success rate of AdaBoost is much higher to classify vulgar comments than Decision Tree and Naïve Bayes Algorithm. The authors shows that toxic comments and cyber bullying began to turn into a significant issue of the World Wide Web from back in late 1990s. Techniques for web-indecent separating dependent on substance examination were additionally investigated to recognize vulgar pages [6].

S. Shubha et al [7] showed a classification framework to review comment implementing Machine Learning Bayes Sentiment Classification (MLBSC) which was proposed in their

paper. The method firstly extract user comments and related comments are listed based on prior training. Then they evaluated label of class applying probabilistic Bayes classifiers. Finally ,they got 88.75% accuracy applying MLBSC on 70 testing comments.

1.3 Thesis Orientation

- **Chapter 2** Background Studies
- **Chapter 3** Proposed Methodology
- **Chapter 4** Result and Analysis
- **Chapter 5** Conclusion and Future works
- **Chapter 6** Reference

Chapter 2

BACKGROUND STUDIES

The following explanations will give all the elaborations and details of all the algorithms and sensors that have been used for this research. Each of the algorithms and sensors play an important role in the crop monitoring system.

2.1Vulgarity in Social Media and Virtual life

Now a days we saw a blast of informal organizations, for example, Facebook, which added another social measurement to the web. There has been a quick rise of online associations between gatherings of individuals who share comparative interests, however they are congregated in an outright space. various sites (e.g., Facebook, Youtube, Twitter, Google in addition to) have actualized dynamic social substance in which online networks can be assembled and continued effectively through the assistance of social associations and interchanges between clients. While such systems have made individuals, networks and gatherings with shared interests remain progressively "associated," risky utilization of Internet and informal organization specifically likewise began being perceived as mental issue everywhere throughout the world. While the primary focal point of studiesin psychiatry relating to web in a decade ago of twentieth century was on Internet "compulsion," it moved to the issue identified with the utilization of informal communication locales in the main decade of 21st century. Youthful grown-ups, especially youngsters would in general be uninformed of exactly how much time they truly spent on long range informal communication destinations, and the impact this may have on their scholarly execution and social connection. It has additionally been noted in concentrates that there might be a relationship between's low confidence and a feeling of social insufficiency and informal organization addiction.From scholarly tarrying to social hindrance to the extent genuine

physical collaborations are concerned, lessened efficiency at work and physical issues related with an inactive way of life; there appear to be sufficient issues identified with Internet and interpersonal organization addictions to give analysts enough to chip away at for a long time to come. Proof of this pattern can be seen by mechanical reports from offices that screen the exercises of online clients. In 2009, it was accounted for that a normal informal organization client around the globe spent more than 5½ h every month on long range interpersonal communication locales, which was triple the time spent on other online exercises, for example, web perusing. From April 2008 to April 2009, the absolute minutes spent on Facebook in U.S., specifically, has expanded from 1.7 billion minutes to 13.9 billion minutes (700% yearly growth).Alongside these figures, examine contemplates demonstrate that amiability of the Internet is in charge of the intemperate measure of time people spend having associations by means of gatherings, web based diversions, and blogs. At the end of the day, the ongoing rise of these new online networking innovations has changed the idea of the Internet just as its use. Accordingly, discoveries from earlier examinations on web conditions, over the top use or addictions may not be legitimate in this new setting. The subjects must utilize web either at home or at different spots like digital bistros. Tricky utilization of long range informal communication destinations portrayed by inordinate extravagance in web person to person communication unfavorably influencing scholarly and co-curricular exercises too asocial and relational conduct was surveyed by youthful's web dependence test altered for risky interpersonal interaction use. Each inquiry in the Young's web enslavement test was solicited explicitly in setting from person to person communication locales.

2.2 Algorithms

For the classification and comparison purpose, we've used some supervised learning algorithm in this thesis. Supervised learning is basically trained on pre-defined data. Supervised learning based algorithms analyses training data and gives function which can be used to calculate output from a certain input.The individuals who don't give wanted output will be determined from a given new information. For instance, to prepare a spam classifier calculation, a preparation dataset of human labeled spam or non-spam messages is utilized and after that utilizes the calculation to order whether another email without tag is spam or not.For content arrangement each archive is leveled by at least zero classifications and new content will be ordered by learning a classifier. Archives which are named will be considered as positive and others as a negative precedent. The errand is to discover a weight vector for a content classifier which arranges new content archives. In this thesis, we discussed algorithms like Convolutional Neural Networks, Support Vector Machines, Random Forest, Decision Tree to compare our model.

2.2.1Support Vector Machine

Support Vector Machine (SVM) [8] is a managed AI algorithmic standard which may be utilized for every arrangement or relapse difficulties and it's basically used in arrangement issues. In this algorithmic standard, we plot every datum thing as a point in n-dimensional space where n is number of highlights one has with the estimation of each component being the estimation of a specific arrange. At that point, we perform order by finding the hyper-plane that separates the two classes well.

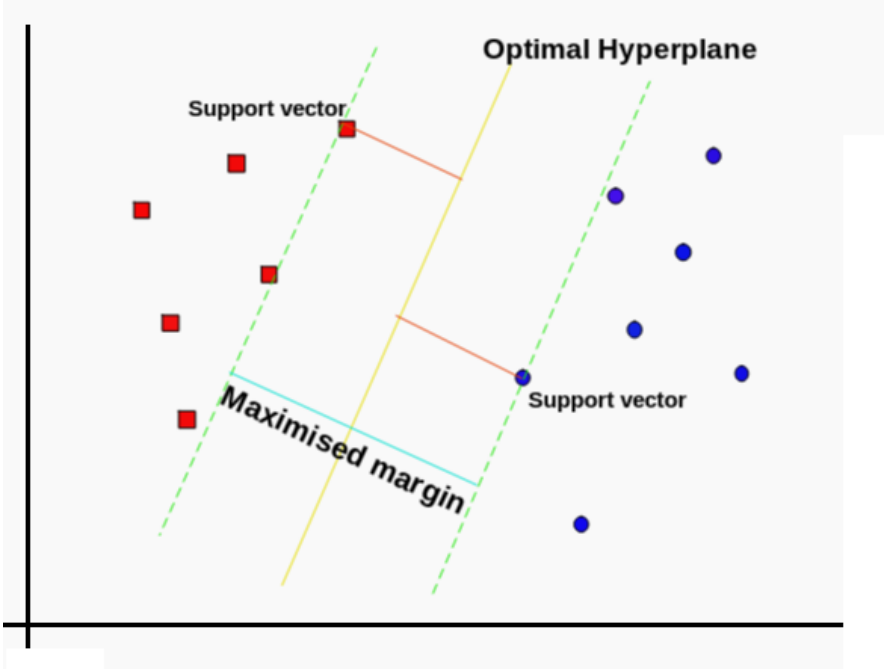


Figure 1. SVM Architecture [9]

Regularly analysts will in general plot each learning thing as some degree in n-dimensional territory with the value of each element having the value of a chosen arrange. At that point, to perform arrangement pointingthe hyper-plane that separate the 2 classes fine. It is a nonprobabilistic double direct classifier, how-ever are regularly controlled amid a way that it will perform non-straight and probabilistic characterization additionally, making it flexible algorithmic program. A SVM model could be an outline of the occasions as focuses in territory mapped, so they will be ordered and partitioned by a straightforward hole. New occasions are then mapped into the indistinguishable region and anticipated that inside which class it would be upheld which part of the hole they fall in. the most points of interest of SVM is that the unquestionable truth that it's powerful in high dimensional zones.

2.2.2Decision Tree Classification

Decision tree [10] might be a learning algorithmwhich is utilized for characterization and regression. It is applied by slicing the data into 2 or extra subsets supportthe estimations of info factors. An esteem work or discordant foundation is utilized to see the best split among all the split focuses. The data is part recursively into groups till the leaves contain only one example. Amid this model, partner degree advanced adaptation of the CART rule is utilized to execute the decision tree classifier. Call trees are clear to decipher and see, contrasted with many arrangement calculations. Also, call trees need next to no preprocessing as anomalies don't affect the execution. Besides, they're not bolstered the Euclidian separation. Henceforth, highlight scaling isn't required. Additionally, include scaling may prompt wrong suspicions being implicit since the qualities would be adjusted. Call trees will deal with each unmitigated and numerical factors as information along these lines it's worthy for this model, since the data set contains every factor assortments. Amid this model the connection between the component variable and target variable is confounded and high non-straight. Consequently a call tree contains a bigger probability of outflanking lin-ear models like arrangement relapse. While call Tree have numerous advantages, they even have a few inconveniences. One is that, bring Trees will cause over fitting by making a tree that is excessively confounded and hence doesn't foresee well on new data. At last, since call Trees are ravenous algorithms, the ideal tree isn't basically returned.

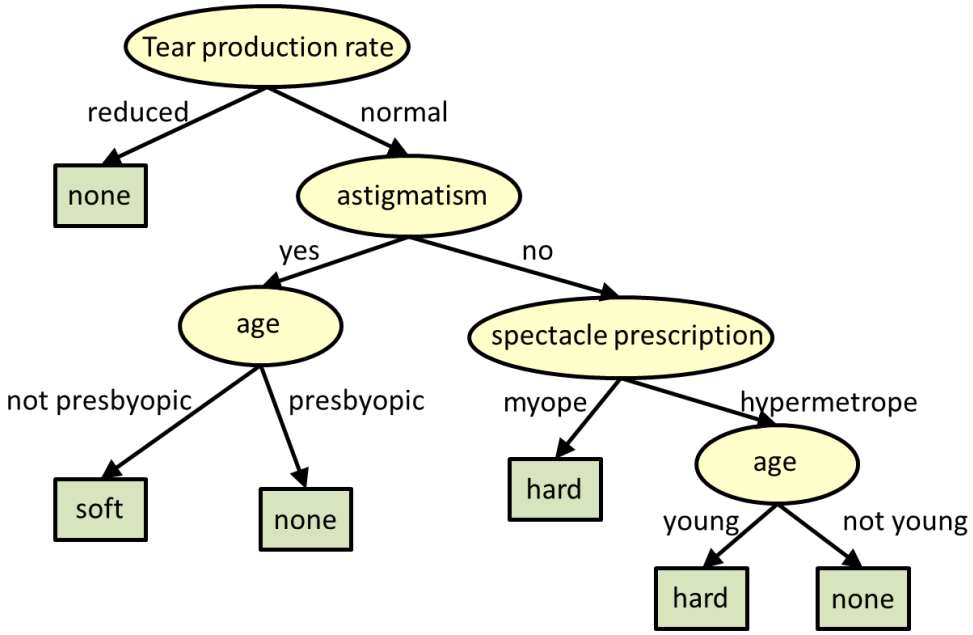


Figure 2. Architecture of Decision Tree Classifier [11]

2.2.3Convolutional Neural Networks (CNN)

Convolutional Neural Networks are multistage trainable Neural Networks architectures developed for classification tasks. Each of these stages, consist the types of layers described below,

Convolutional Layers: convolutional layers are core element of the CNNs. A convolutional layer contain kernel matrices that work convolution on their given input and produce an output matrix of features where a bias is added. The learning procedures aim to train the kernel weights and biases as shared neuron connection weights. Training biases for sharing neuron connection weights and the kernel weights.

Pooling Layers: Pooling layers are likewise indispensable parts of the CNNs. The motivation behind a pooling layer is to perform dimensionality decrease of the information include pictures. Pooling layers make a subsampling to the yield of the convolutional layer frameworks brushing neighboring components. The most widely recognized pooling capacity is the maximum pooling capacity, which takes the greatest estimation of the nearby neighborhoods.

Chapter 3

PROPOSED METHDOLOGY

The proposed methodology is divided into two core part CNN for text classification with feature extraction implementing natural language processing. Figure 1 shows the work diagram of the methodology.

Figure 3. Proposed work diagram

3.1 Data Acquisition

We have used two types of dataset in this research, one is acquired from Kaggle which is very popular publicly available dataset named “Wikipedia Talk Page Comments annotated with toxicity reasons” [ ] which content almost 1,60,000 comments with manually labelling and another one is our own made dataset extracting comments from YouTube, Facebook and tweeter for testing only.

The dataset contains total six classes which are described down below,

a. **Toxic** : a common class for all comments which are vulgar, toxic and bullying type.
b. **Sever toxic :**This class denotes the extreme toxic or vulgar comments.
c. **Threat :**Articulations with the aim to cause antagonistic activity .
d. **Obscene :**Rude, dirty, wicked types of languages are in this category.
e. **Identity :**Racism, sexism, bullying for homophobia etc are in this category.
f. **Insult :** Hate speech, verbal abuse and showing disrespect, commenting aiming the disability or religion of specific people .

In the training dataset, these six categories are binarily individually labelled where one comment or sentence can belong to more than one category or no category if they are not toxic or vulgar. Figure 3. shows the training dataset before preprocessing.

id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
0000997932d777	Explanation	0	0	0	0	0	0
000103f0d9cfb60f	D'aww! He matches this background colour I'm seemingly stu	0	0	0	0	0	0
000113f07ec002fc	Hey man, I'm really not trying to edit war. It's just that this guy i	0	0	0	0	0	0
0001b41b1c6bb37	"	0	0	0	0	0	0
0001d958c54c6e3	You, sir, are my hero. Any chance you remember what page	0	0	0	0	0	0
00025465d4725e	"	0	0	0	0	0	0
0002bcb3da6cb3	COCKSUCKER BEFORE YOU PISS AROUND ON MY WORK	1	1	1	0	1	0
00031b1e95af792	Your vandalism to the Matt Shirvington article has been reve	0	0	0	0	0	0
0003726f1536c51	Sorry if the word 'nonsense' was offensive to you. Anyway, I'r	0	0	0	0	0	0
00040093b2687c	alignment on this subject and which are contrary to those of	0	0	0	0	0	0
0005300084f90ec	"	0	0	0	0	0	0
00054a5e18b50dc	bbq	0	0	0	0	0	0
0005c987bdfc9d4	Hey... what is it..	1	0	0	0	0	0
0006f16e4e9f292	Before you start throwing accusations and warnings at me,	0	0	0	0	0	0
00070ef96486d6f	Oh, and the girl above started her arguments with me. She st	0	0	0	0	0	0
00078f8ce7eb27f	"	0	0	0	0	0	0
0007e25b2121310	Bye!	1	0	0	0	0	0
000897889268bc	REDIRECT Talk:Voydan Pop Georgiev- Chernodrinski	0	0	0	0	0	0
0009801bd85e58	The Mitsurugi point made no sense - why not argue to includ	0	0	0	0	0	0
0009eaea3325de	Don't mean to bother you	0	0	0	0	0	0
000b08c464718	"	0	0	0	0	0	0
000bfd0867774	"	0	0	0	0	0	0
000c0dfd3958	"	0	0	0	0	0	0
000c6a3f0cd3ba	"	0	0	0	0	0	0
000cfee90f50d47	"	0	0	0	0	0	0
000eefc67a2c93	Radial symmetry	0	0	0	0	0	0
000f35deef84dc4	There's no need to apologize. A Wikipedia article is made for	0	0	0	0	0	0

85e1aab1df677bf4	Please do not add nonsense to Wikipedia. It is considered va	0	0	0	0	0	0
85e25e1f3926bc9	"After reading the info on Rex Humbard, I noticed not ONE	0	0	0	0	0	0
85e3b61f4872bb0	"It is somewhat blurry... although I never really thought that a	0	0	0	0	0	0
85e4917d329292(	a QUESTION	0	0	0	0	0	0
85e8727a493f48(	I added Spain as a country bordering Morroco. Indeed, Morr	0	0	0	0	0	0
85eb4ccdb80defc	Deconstructing a few paragraphs from this source, here's	0	0	0	0	0	0
85eb560f20c632f	Truth be told, the Calvocaressi et al. is a bit ambiguous in pla	0	0	0	0	0	0
85ec814c850bd6f	"	1	0	1	0	1	0
85ede7fc2f01dd0	REDIRECT Talk:Bio Senshi Dan: Increaser to no Tatakai	0	0	0	0	0	0
85ef48669590b53	"	0	0	0	0	0	0
85f02a76e382abf	What is wrong with you?	0	0	0	0	0	0
85f06612b383533	New External Link?	0	0	0	0	0	0
85f169c64a7a374	HEY DUMB FUCK	1	1	1	0	1	0
85f44751a87fe0d3	Does anyone have any suggest as to what this list should/co	0	0	0	0	0	0
85f4d22d47c9373	I guess I need to go restart my modem and get a new IP addre	0	0	0	0	0	0
85f72d361e6d8e5	"==selected anniversary==	0	0	0	0	0	0
85f7b91d8d7aec9	14th Dalai Lama ==	0	0	0	0	0	0
85f933862693764	"	0	0	0	0	0	0
85f9669eb8f096c	block	0	0	0	0	0	0
85fb01552647828	The main contributor to this article changed the date of the	0	0	0	0	0	0
86004958423e23	"	0	0	0	0	0	0
8601ad709ca455!	Independent proposal for WP:CAL and WP:SOCAL tags	0	0	0	0	0	0
8603bea1881ba95	"	0	0	0	0	0	0
860827556955f12	A great truth	1	0	1	0	1	0
8609ff83be3a8ab	Huh? When I put that comparison on the Photoshop article it	0	0	0	0	0	0
860d4e2ad1a21b4	, 17 December 2011 (UTC)	0	0	0	0	0	0
860de3bd26cc12a	"I agree, we need to have the article unlocked so we can fix	0	0	0	0	0	0
860f7326ebb8722	"	0	0	0	0	0	0
860fe5d33cefab6	Please stop reverting factual edits	0	0	0	0	0	0

Figure 4. Labelled training dataset

3.2 Data Pre-processing

In preprocessing, first of all null data are removed so that it may not occur any trouble while the processing portion. A blank row of comment is counted as a null data and is removed. After that, the huge amount of data needs to be separated into vulgar and non vulgar part. To do this, we implement sampling technical, which simply search is every comment if there is class with value 1 which indicates it is in any of vulgar class. If at least one class is true, then it is vulgar, rest are non vulgar.

The sampled dataset are stored in vulgar and non vulgar part which are shown in figure 4 and 5 and these will be processed separately in feature extraction, tokenization . Figure 6 shows the vulgar content occurrence in the whole dataset.

	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
6	0002bcb3da6cb337	COCKSUCKER BEFORE YOU PISS AROUND ON MY WORK	1	1	1	0	1	0
12	0005c987bdfc9d4b	Hey... what is it..\n@ | talk .\nWhat is it.....	1	0	0	0	0	0
16	0007e25b2121310b	Bye! \n\nDon't look, come or think of comming ...	1	0	0	0	0	0
42	001810bf8c45bf5f	You are gay or antisemmitian? \n\nArchangel WH...	1	0	1	0	1	1
43	00190820581d90ce	FUCK YOUR FILTHY MOTHER IN THE ASS, DRY!	1	0	1	0	1	0

Figure 5. Sampled vulgar comment

3.3 Feature Analysis

We In order to differentiate between vulgar and legitimate comments, we compute nine features for each comment. The features are assigned numeric values and they represent various differentiating characteristics of the two types of comments. The features are as follows:

1. *URL in the replies:* harassing comments may get an increment of responses regardless of the fact that spammers need to divert users to sites with a relatively small number of viewers in order to maximize the ranking of the website.That is the reason the quantity of connections in the remarks must be taken into records when distinguishing toxic comment.

2. *Spaces in the comments:* Vulgar comments may have lots of white spaces beacause replies or comments creates huge impact on the user who reads these comments.

3. *Sentences in the comment:* Total sentences in a toxic comment is lower than the total sentences in a rational statement despite the fact that use of words and sentences is generally consistent.

4. *Punctuation marks :* Presence of large number of punctuation marks, especially exclamation and dot characters is a defining feature of vulgar comments, as punctuation characters tend to attract the attention of the reader. In contrast, legitimate comments tend to contain only sentence delimiting punctuation marks. However, sometimes legitimate comments can contain larger than normal number of punctuation characters to express strong feelings.

5. Word duplication: Vulgar or toxic comments tend to have a lot of repetition of words compared to legitimate comments with contextual flow of word structure. To capture this property we take the ratio of the number of unique words to the total word count in the comment and define it as word duplication ratio RWD.The equation is expressed in the following way:

$$RWD = \frac{Number\ \ of\ Unique\ \ words\ \ in\ a\ comment}{Total\ \ numbers\ \ of\ \ words\ \ in\ a\ comment} \qquad (1)$$

6. *Ratio of Stop Words:* Another feature that captures the difference between vulgar comments and the legitimate comments is the ratio of stop word, which is computed as the ratio of the number of stop words to the number of total word count in the comment.

$$RSW = \frac{Number\ \ of\ Stop\ \ words\ \ in\ a\ comment}{Total\ \ numbers\ \ of\ \ words\ \ in\ a\ comment} \qquad (2)$$

7. Non-ASCII Characters: Vulgar comments may contain a moderate amount of Non-ASCII Characters compared to legitimate comments and non-toxic comments.

8. Capital Letters: People who write vulgar comments, use more capital letters than legitimate . Sometimes, vulgar or toxic comments are very often written fully using capital letters.

9. New lines numbers: Vulgar comments contain a huge number of new lines in order to draw attention with the facts by creating a visual effect for the reader.

3.4 Gluon Natural Language Processing

Nature Language Processing (NLP) is a hypothesis propelled scope of computational procedures for the programmed analysis and portrayal of human language. Practically, It is a procedure which empowers a machine to process a characteristic language (like English) and perform all of the things that a human can do. Prior to going into profound ideas of NLP, a lot of deficient sentences which regularly shows up in a tweet, youtube and facebook comment are recognized. Subsequent to looking into on dataset, some regular sentences in spam remarked were found. They are 'include me at', 'take me out on the town', 'you'll giggle when you see this pic of you', 'You appear to be unique in this photograph', 'my companion sent me this pic with you in it', 'my companion demonstrated to me this pic of you', 'tail me back', 'markdown drugs', 'I discovered you in this video', and some other sensitive words that can't be written for unable circumstances. If these expressions are found in the facebook, email, youtube then the user is classified as spam. In this thesis, three elements of NLP are applied which are (1) stop words removing ,(2) tokenization and (3) stemming. For processing English there is no need of stop words like I, you, in, the, we ,is, was, me, our,it, etc. o each of these words are expelled and just the keyhwords are removed. The following stage is to discover the root word or stem of the catchphrase. For this, stemming methods are utilized. A basic stemming calculation has been utilized in this paper. A lot of spam words that can show up in a tweet is distinguished, similar to 'pornography', 'Viagra' and so on. The stemmed watchwords are contrasted and the arrangement of recognized spam words. On the off chance that the words coordinate, at that point the client is viewed as spam. At this stage, in the event

that the client isn't found as spam, at that point the third system of Machine Learning is utilized.

Gluon comes with a really propelled NLP toolbox, which makes working with content simple. It additionally fuses pre-prepared Language Models, the mystery sauce for Transfer Learning. Thing is, how might we anticipate that a model should make sense of whether content is lethal or not, on the off chance that it can't "communicate in" English by any means. To implement Gluon with NLP, first of all, an independent language model is trained before embedding matrix by a standard LSTM encoder. Then LSTM output is pooled which is generated by sequence of tokenization aiming to feed dense layer.

3.4.1 Stop Word Removing

Stop word expulsion is a standout amongst the most usually utilized preparing strategy at the present time. Web indexes utilized this procedure to disregard each one of those words or character that has no esteem or less incentive in the season of creating precision structure the dataset. At the season of making the list, most motors are modified to evacuate some specific words that has less weight. Chiefly the rundown of words or characters that are not added to the last informational index is known as the stop word list. This recoveries both existence as it evacuates the words at ordering time and overlooked at looking time. We expel these arrangement of words as they convey no data for example pronoun, relational word, conjunctions. In English language there are more than 500 stop words [23]. These are a few instances of stop words which encourages us to manufacture the framework.

3.4.2 Tokenization

In natural language processing ,tokenization is applied for word segmentation. To effectively make an interpretation of this unstructured text data into machine-interpretable data, we separate pieces of constant content information into a rundown of words, at that point encode

them into numerical vectors. We at that point encode every interesting word to its numerical portrayal. As appeared in Figure 5, mapping this tokenization to sectioned content information basically restores the content as a rundown of numeric components, which speak to included words in the vocabulary. This concise clarification serves to give establishments to the "word2vec" method is applied.

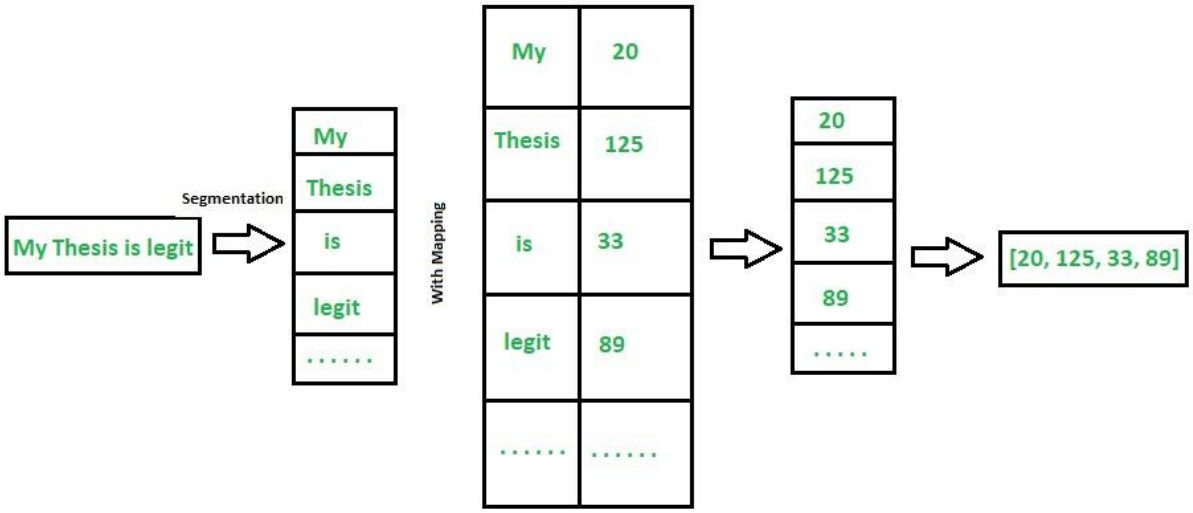


Figure 6. Process of tokenization

Word2Vec is developed by a team of google back in 2013 which is a combination of models to create word embedding which retained context of words[ ]. Word2vec models are shallow, two-layer neural systems built dependent on the possibility that comparable words would show up in comparable positions in the specific circumstance.

Vectors are calculated by the cosine angle between them with following equation,

$$\cos\theta = \frac{x.y}{|x||y|} \qquad (3)$$

Large dot product $x.y = \cos\theta\,|x|||y|$ denotes more similarity. Word2vec presented two methodologies in figuring the word implanting with the goal that comparable word vectors have higher speck item.

3.4.3 Stemming

Stemming is the procedure to diminish a group of comparable words and give them a root word. Stemming is including word and giving a lot of words same sort of importance [23]. Here root word is called lemma. Stemming is essential in characteristic language preparing (NLP). In a major informational collection a great deal of word can be comparable, in the event that we can evacuate the comparative significance words the span of our dataset will diminish and it will lessen run time. When we found a comparative kind of word we give a specific load to it for estimating [23]. At the point when another word is discovered it is chosen as another kind. To get the best outcome the procedure of lemma is utilized. To discover lemma, stemming is performed by various AI calculations. A few calculations will basically strip perceived prefix and postfix. Be that as it may, these sort of execution makes a great deal of blunder which isn't appropriate for learning calculations. For a model 'responsively' can be decreased to a word like 'respon'. In addition stripping the calculations realizes the postfix words and expels the addition. Another is Lemmatization. Here the calculation separates the word to its root catalog. At that point it arrange the word type as indicated by standards of language structure. Another sort is Stochastic, it gets the type of how the root words could be arched and it expels the contaminated types of words.

3.5 Post Similarity Analysis

The initial step the module performs is to analysis the comment by applying the accompanying activities: stop word evacuation and tokenization. In the subsequent stage the

comment is parsed at token dimension so as to recover a list of equivalent words for each word from the comment. So as to distinguish a rundown of equivalent words for a given word from the comment we executed the equivalent word recovery submodule that makes an association with an online lexicon Thesaurus.com and makes a solicitation to parse the source page and to remove every one of the equivalent words for a particular word. The rundown of equivalent words is recovered and it is put away inside the post-remark similitude module. For each word inside the remark we seek if a similar word shows up in the related post or if equivalent words of the word show up in the related post and we figure the aggregate of the frequencies of events of each word and its equivalent words from the remark in the post. The recipe is given by the standardized estimation of the total of the frequencies of events of each word and its equivalent words from the comment in the post.

3.6 Text Classification using CNN

Convolutional Neural Networksare multistage trainable Neural Networks architectures developed for classification tasks. Each of these stages, consist the types of layers described below [7, 8]:

3.5.1 *Convolutional Layers:* convolutional layers are core element of the CNNs. A convolutional layer contain kernel matrices that work convolution on their given input and produce an output matrix of features where a bias is added. The learning procedures aim to train the kernel weights and biases as shared neuron connection weights. Training biases for sharing neuron connection weights and the kernel weights.

3.6.2 *Pooling Layers:* Pooling layers are likewise indispensable parts of the CNNs. The motivation behind a pooling layer is to perform dimensionality decrease of the information include pictures. Pooling layers make a subsampling to the yield of the convolutional layer

frameworks brushing neighboring components. The most widely recognized pooling capacity is the maximum pooling capacity, which takes the greatest estimation of the nearby neighborhoods.

3.6.3*Fully-Connected Layer:* Fully connected layeris a great Feed-Forward Neural Network (FNN) shrouded layer. It very well may be translated as an uncommon instance of the convolutional layer with portion estimate 1 × 1. This sort of layer has a place with the class of trainable layer loads and it is utilized in the last phases of CNNs.

3.6.4*Embedding Layer:* Embedding Layer is an important side of the CNNs for substance portrayal issues. The inspiration driving an embedding layer is to change the substance commitments to a proper structure for the CNN. Here, every statement of a substance report is changed into a thick vector of fixed size.

Convolutional Neural Networks is mainly popular for image classification because of the abily to exploit 2 statistical characters which are local stationary and compositional structure. In this thesis, CNN is applied for text classification where main dataset represents the mentioned two statistical characters relying on the fact that neighboring words in a sentence present dependency, however, their processing is not straight forward. In image classification, pixels are some integer values with specific threshold value, but in the case of sentence or words, we need to encode first before fed to the networks [ ]. To do so, we applied NLTK library [ ]of python for using vocabulary which is structured as an list containing words which are shown in the set of comment's texts. Then we need to map each word to encode it integer ranging between one to size of vocabulary.

The fluctuation in reports length (number of words in a record) should be tended to as CNNs require a consistent information dimensionality. For this reason the cushioning strategy is received, loading up with zeros the archive grid so as to achieve the most extreme length

among all reports in dimensionality. In the subsequent stage the encoded reports are changed into lattices for which each column compares to single word. The created lattices go through the implanting layer where each word (push) is changed into a low-measurement portrayal by a thick vector [11]. When all is said in done the word installing strategies have been prepared on a substantial volume dataset of words delivering for each word a thick vector with a particular measurement and fixed qualities. The word2vec implanting strategy for instance, has been prepared on 100 billion words from Google News delivering a vocabulary of 3 million words. The inserting layer coordinates the info words with the fixed thick vector of the pre-prepared implanting techniques that have been chosen. The estimations of these vectors don't change amid the preparation procedure, except if there are words not officially incorporated into the vocabulary of the inserting strategy in which case they are instated arbitrarily.

3.7 Evaluations

To evaluate the model, we have computed recall, f1 score, accuracy, sensitivity, precision to compare the model with each algorithm. Figure 6 shows the confusion matrix.

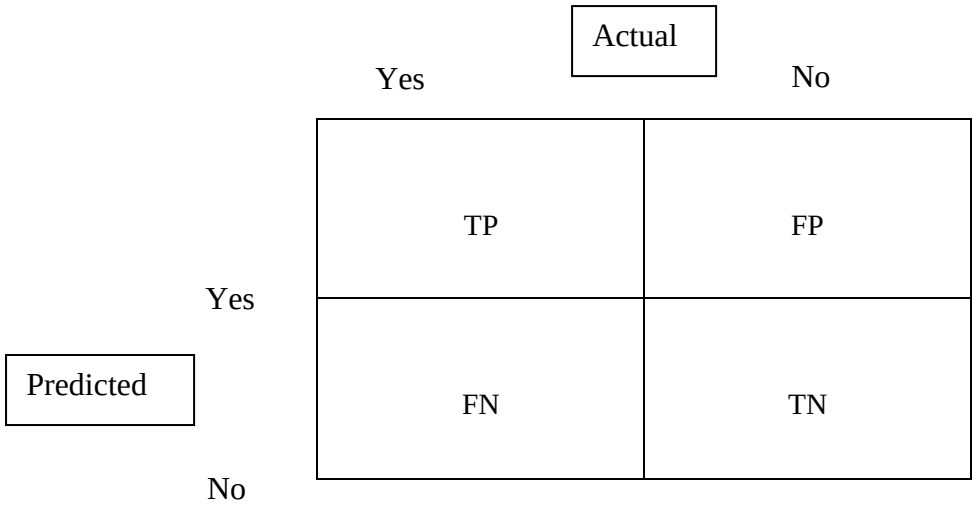


Figure 7. Confusion Matrix

Where,

TP = True Positive

FP = False Positive

TN = True Negative

FN = False Negative

Using the following equation, recall, f1 score, accuracy, sensitivity, precision are calculated,

$$\left.\begin{aligned} Recall &= \frac{TP}{TP+FN} \\ Precision &= \frac{TP}{TP+FP} \\ Accuracy &= \frac{TP+TN}{TP+FP+TN+FN} \\ Sentivity &= \frac{TP}{TP+FN} \end{aligned}\right\} \quad (4)$$

Chapter 4

RESULT AND ANALYSIS

Implementing the dataset from Kaggle [19] and extract commenting from youtube, facebook and tweeter using chrome's comment extraction extenstion, almost 1,59,000 data are stored intraining dataset with labeling . The relation between dataset and classes are shown below

	toxic	severe_toxic	obscene	threat	insult	identity_hate	other
toxic	1	0.308619	0.676515	0.157058	0.647518	0.266009	-0.967748
severe_toxic	0.308619	1	0.403014	0.123601	0.375807	0.2016	-0.298666
obscene	0.676515	0.403014	1	0.141179	0.741272	0.286867	-0.702812
threat	0.157058	0.123601	0.141179	1	0.150022	0.115128	-0.162925
insult	0.647518	0.375807	0.741272	0.150022	1	0.337736	-0.677324
identity_hate	0.266009	0.2016	0.286867	0.115128	0.337736	1	-0.280144
other	-0.967748	-0.298666	-0.702812	-0.162925	-0.677324	-0.280144	1

for visualization:

Figure 8. Relation between labeled dataset

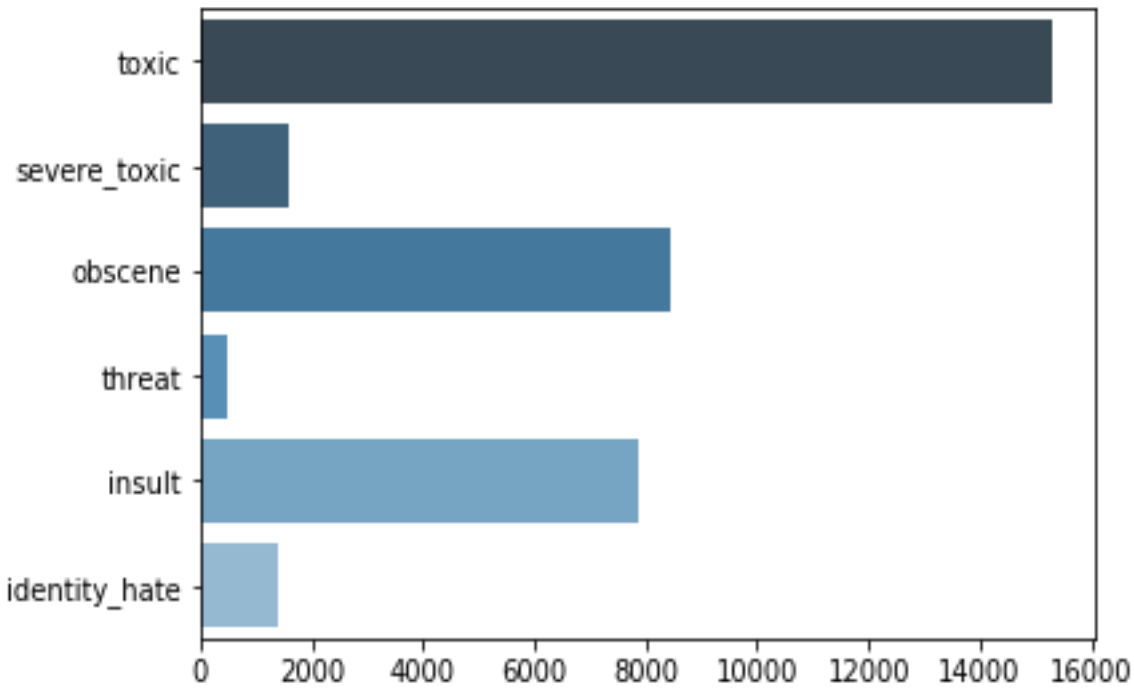
The dataset is labelled based on toxic, severe toxic, Obsecene, threat, insult, identity hate speech and others; where 1 means presence of its class, 0 mean not in that class. A comment or text can belong to multiple classes. Figure 8 shows relation of each classes of the labelled dataset .

```
Train on 127656 samples, validate on 31915 samples
Epoch 1/5
127656/127656 [==============================] - 467s 4ms/step - loss: 0.3026 - acc: 0.9938 - val_loss: 0.3034 - val_acc: 0.994
1
Epoch 2/5
127656/127656 [==============================] - 447s 4ms/step - loss: 0.2990 - acc: 0.9939 - val_loss: 0.3028 - val_acc: 0.993
9
Epoch 3/5
127656/127656 [==============================] - 447s 3ms/step - loss: 0.2977 - acc: 0.9937 - val_loss: 0.3019 - val_acc: 0.993
7
Epoch 4/5
127656/127656 [==============================] - 453s 4ms/step - loss: 0.2962 - acc: 0.9935 - val_loss: 0.3008 - val_acc: 0.993
5
Epoch 5/5
127656/127656 [==============================] - 436s 3ms/step - loss: 0.2943 - acc: 0.9932 - val_loss: 0.3004 - val_acc: 0.993
2
yoo
```

(a)

```
array([[0.6196143 , 0.06325791, 0.3149264 , 0.02090574, 0.3456367 ,
        0.09032118],
       [0.6307527 , 0.0465546 , 0.2922022 , 0.01507705, 0.32081485,
        0.06480588],
       [0.62810767, 0.04707333, 0.29225576, 0.01507414, 0.32434827,
        0.06559952],
       [0.65722156, 0.03800321, 0.28731743, 0.01173671, 0.299912  ,
        0.05303184],
       [0.6308709 , 0.04626049, 0.29339245, 0.01473308, 0.32179758,
        0.06375712],
       [0.6440422 , 0.03917455, 0.28352523, 0.01280564, 0.31083292,
        0.05863864],
       [0.66531056, 0.03335576, 0.28692177, 0.0099727 , 0.29864153,
        0.04750201],
       [0.6304702 , 0.04552723, 0.28712666, 0.0151311 , 0.3216888 ,
        0.06469539],
       [0.6705599 , 0.03378842, 0.2675668 , 0.01310985, 0.28182095,
        0.05308157],
       [0.65492046, 0.03537232, 0.27905625, 0.01188434, 0.3053456 ,
        0.05498387]], dtype=float32)
```

(b)



(c)

Figure 8. (a,b) Working Procedure and output of CNN, (c)Final classified result visualization

The CNN classifier classifies based on extracting the eight features and implementing gluon natural language processing where it rejects all unnecessary words by stopwords and tokenization, removing common sentence structures, numbers, barriers of url and so on.After the post comment analysis, a list of vulgar words is processing with NLP to classy the vulgar words from test dataset. Figure 9 shows the output of the system.

Figure 5. displayed the multiclass output of classification to visualize where toxic levelis in more than 15,000 comments, obseene is in more than 8,500 comments and so on.The architecture is got 95.4% accuracy with CNN and NLP. Table I shows the comparison of the model with other classifying algorithms where our proposed models gives the best accuracy.

Table I. Result comparison with other Classifiers based on Accuracy

Classifier	*Accuracy*
Random Forest	94%
SupportVector Machine	91%
Decision Tree Classification	89.75%
Proposed Model	95.4%

Table II. Result comparison with other Classifiers based on Sensitivity

Classifier	*Sensitivity*
Random Forest	86.08%
SupportVector Machine	82.55%
Decision Tree Classification	71.95%
Proposed Model	82.32%

Table III. Result comparison with other Classifiers based on F1 Score

Classifier	**F1 Score**
Random Forest	91.07%
SupportVector Machine	86.87%
Decision Tree Classification	83.45%
Proposed Model	95.4%

Chapter 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

We led an inside and out examination on remarks to reveal some insight into various highlights of spam remarks. We were keen on structure a framework which distinguishes vulgar or spam and non-vulgar comments as for various attributes that are characterized. The combination of Gluon NLP with tokenization, stop words, stemming and nine numerical features and huge dataset made the result more error free. In our characterization tests, we showed that our usage of the spam discovery framework gives the best outcomes by utilizing the Gluon NLP and CNN classifier where we got 95.4% accuracy, 82.32% sensitivity and 95.4% f1 score which are higher than other regular algorithms.

5.2 Future Work

In our future work, we need to acquaint another module with our spam identification framework. This module will be utilized so as to perform estimation investigation on spam and authentic remarks. We need to make an examination between the opinion scores acquired for spam remarks and the notion scores got for authentic remarks. The conclusion score can be additionally utilized as an info highlight for the classifier which we constructed. The motivation behind why we trust the classifier will give better outcomes by mulling over the suppositions transmitted by the essayist in the remark is on the grounds that in our examination we saw that spam remarks typically will in general express increasingly negative emotions and real comment.

References

[1] V. Singh and A. K. Misra, “Detection of plant leaf diseases using image segmentation and soft computing techniques,” Information Processing in Agriculture, vol. 4, no. 1, pp. 41–49, 2017.

[2] S. Srivastava, P. Khurana, V. Tewari,”Identifying Aggression and Toxicity in Comments using Capsule Network.” In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), pp. 98-105, 2018.

[3] C. Rădulescu, M. Dinsoreanu and R. Potolea, "Identification of spam comments using natural language processing techniques," 2014 IEEE 10th International Conference on Intelligent Computer Communication and Processing (ICCP), Cluj Napoca, 2014, pp. 29-35.

[4] M. Mccord and M. Chuah, “Spam Detection on Twitter Using Traditional Classifiers,” Lecture Notes in Computer Science Autonomic and Trusted Computing, pp. 175–186, 2011.

[5] K. Kandasamy and P. Koroth, “An integrated approach to spam classification on Twitter using URL analysis, natural language processing and machine learning techniques,” 2014 IEEE Students Conference on Electrical, Electronics and Computer Science, 2014.

[6] X. Carreras, L. Marquez, “Boosting trees for anti-spam email filtering”. In Proceedings of RANLP-01, 4th International Conference on Recent Advances in Natural Language Processing, 2001, pp. 58-64.

[7] S. Shubha and P. Suresh, "An efficient machine Learning Bayes Sentiment Classification method based on review comments," 2017 IEEE International Conference on Current Trends in Advanced Computing (ICCTAC), Bangalore, 2017, pp. 1-6. [8] B.D. Davison,

“Recognizing Nepotistic Links on the Web”. In AAAI 2000 Workshop on Artificial Intelligence for Web Search, pp.23- 28, 2000.

[9] S. Patil, A. Gune and M. Nene, "Convolutional neural networks for text categorization with latent semantic analysis," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, 2017, pp. 499-503. doi: 10.1109/ICECDS.2017.8390217

[10] L. Li, L. Xiao, N. Wang, G. Yang and J. Zhang, "Text classification method based on convolution neural network," 2017 3rd IEEE International Conference on Computer and Communications (ICCC), Chengdu, 2017, pp. 1985-1989. doi: 10.1109/CompComm.2017.8322884

[11] R. Ju, P. Zhou, C. H. Li and L. Liu, “An Efficient Method for Document Categorization Based on Word2vec and Latent Semantic Analysis”, CIT/IUCC/DASC/PICom, pp. 2276-2283, IEEE, 2015

[12] Y. Zhang, By. Wallace, “A Sensitivity Analysis of (and Practitioners’ Guide to) Convolutional Neural Networks for Sentence Classification", arXiv:1510.03820 “Toxic Comment Classification Challenge,” Kaggle. [Online]. Available: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data. [Accessed: 30-Jan-2018].

[13] S. R. Seyyed, M. Fakhrahmad and M. H. Sadredini, "PTokenizer: POS tagger Tokenizer," 2015 2nd International Conference on Knowledge-Based Engineering and Innovation (KBEI), Tehran, 2015, pp. 252-256.

[14] S. Popova, L. Kovriguina, D. Mouromtsev and I. Khodyrev, "Stop-words in keyphrase extraction problem," 14th Conference of Open Innovation Association FRUCT, Espoo, 2013, pp. 113-121.

[15] “Natural Language Toolkit¶,” Natural Language Toolkit - NLTK 3.4.1 documentation. [Online]. Available: https://www.nltk.org/. [Accessed: 15-Jun-2018].

[16] S. Gadri and A. Moussaoui, "Information retrieval: A new multilingual stemmer based on a statistical approach," 2015 3rd International Conference on Control, Engineering & Information Technology (CEIT), Tlemcen, 2015, pp. 1-6.