International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 08, 2020
ISSN: 1475-7192

Detecting Hate Speech on Social Media Using Machine Learning

[1]Aswathy K Cherian, [2]Aakash Tripathi, [3]Shrey

Abstract--*As times have progressed, the usage of social media has exponentially increased. Private and public opinions about a wide assortment of subjects are communicated and spread ceaselessly by means of various online social media platforms. Twitter is one of such platforms that has gained a lot of popularity. Twitter offers organizations and individual users a fast and effective way to advertise and communicate their ideas and thoughts without much hassle. Thus, analyzing customers' perspectives toward day to day events is crucial to success in the market place. Building up a program for notion examination is a way to deal with be utilized to computationally gauge people's perceptions. This project applies sentiment analysis to a dataset containing thousands of tweets relating to a given string that is searched, all using R libraries. Searched strings could include hash tags, usernames, specific words etc. Using the processed output, we are able to determine the sentiments of people regarding any trending topic. Tweets are extracted using R and the data is wrangled by removing emoticons and URLs. Lexical Analysis is used to predict the meaning of tweets and subsequently infer the opinion graphically through ggplots, histogram, pie chart and tables.*

Key words –*Detection, Clustering, Classification, Sentiments, Tweets*

I. INTRODUCTION

A study on wellbeing and security in online networking has developed considerably in the most recent decade. An especially applicable part of this work is recognizing what's more, averting the utilization of different types of oppressive language in online journals, smaller scale websites, and informal communities. Various late examinations have been distributed on this topic, for example, the work by Xu et al. (2012) on recognizing digital tormenting, the recognition of detest discourse (Burnap and Williams ,2015) which was the subject of an ongoing overview (Schmidt and Wiegand, 2017), and the recognition of prejudice (Tulkens et al., 2016) in client created content. The developing enthusiasm for this point inside the examination network is confirm by a few related investigations exhibited in Section 2 and by two later workshops: Text Analytics for Cyber security and Online Safety (TA-COS)1 held in 2016 at LREC what's more, Abusive Language Workshop (AWL)2 held in 2017 at ACL.

In this paper we address the issue of abhor discourse identification utilizing a dataset which contains English tweets explained with three marks:

(1) Abhor discourse (HATE);
(2) Hostile language however no abhor discourse (OFFENSIVE); and

[1]*Department of Computer Science, Engineering, SRM IST, KTR, INDIA,*
[2]*Department of Computer Science, Engineering, SRM IST, KTR, INDIA, aakashtripathi_tripathi@srmuniv.edu.in*
[3]*Department of Computer Science, Engineering, SRM IST, KTR, INDIA, shrey_shardindu@srmuniv.edu.in*

DOI: 10.37200/IJPR/V24I8/PR280115
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

(3) No hostile substance (OK). Most examinations on injurious language up until now (Burnap and Williams, 2015; Djuric et al., 2015; Nobata et al., 2016) have been displayed as parallel order with just a single positive furthermore, one negative classes (for example abhor discourse versus nonhate discourse).

As substantiated by Dinakar et al. (2011), frameworks prepared on such information frequently depend on the recurrence of hostile or non-socially worthy words to recognize the two classes. Dinakar et al. (2011) stress that sometimes "not presence of antagonism or irreverence can mislead the classifier".

In the previous multi decade, there has been an exponential flood in the online action of individuals over the globe. The volume of posts that are made on the web each second keeps running into millions. To add to this, the ascent of web based life stages has prompted flooding to content on the web.

Web based life isn't just a phase where people banter with each other, yet it has ended up being outstandingly huge and fills significantly more needs.

II. METHODS

This paper proposes the detection of sentimental terms that classify the tweets as positive, negative and neutral tweets. We use R Studio with required libraries and Twitter Authentication to access API for the analysis.

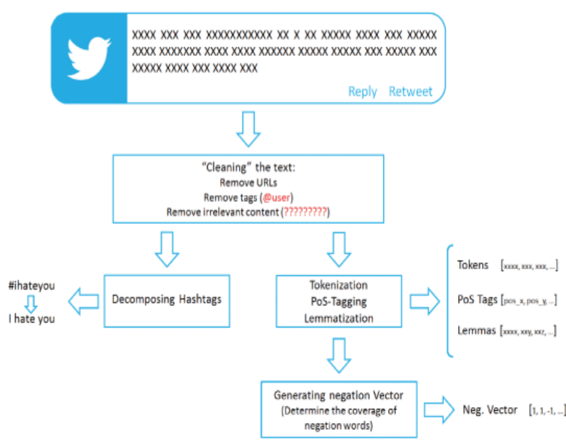

Figure 1: Process of extraction of tweets.

i. Extraction of Tweets

(i) Create twitter application for consumer and access key generation.

(ii) twitteR - Gives an interface to the Twitter web API

(iii) ROAuth - R Interface for OAuth

(iv) Create twitter verified credential object. It is done using consumer key, consumer secret, access token, secret.

(v) During verification, we are redirected to a URL automatically where we click on Authorize app as shown in the image below and enter the unique 7-digit number to get linked to the account from which feeds are being taken.

DOI: 10.37200/IJPR/V24I8/PR280115
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

ii. Data Preprocessing

In an initial stride, we tidy up the tweets. This comprises the expulsion of URLs (that is either beginning with "http://" or "https://") or labels (i.e., "@user").

The next step includes the tokenization, Part-of-Speech (PoS) Tagging, and the vocabulary and morphological analysis of words (using both tokens and PoS tags) of different words. For this purpose, we utilized OpenNLP5 to play out the Natural Language Processing (NLP) undertakings of tokenization and vocabulary and morphological analysis of words.

Afterwards, we produce what we qualify as invalidation vector: we identify the situation of refutation words (e.g., "not," "never," and so on.) and recognize the inclusion of these words. The methodology we utilized is very basic and propelled from crafted by Das and Chen [21]: essentially, an invalidation term includes all the terms that tails it ‘til the following accentuation mark or the event of a differentiation term (e.g., "yet," "in any case," and so on). Terms secured by an invalidation term are given a refutation score equivalent to −1 while the remainder of the terms will be given a score equivalent to 1. This will be utilized in the future to check for positive and negative terms: a positive term (negative term) having an invalidation score equivalent to −1 will be considered as a negative term (positive term), and they are ascribed something contrary to its unique score

iii. Features Extraction

Predominantly 4 arrangements of features are separated which we categorize as "sentiment-based features", "semantic features," "unigram features", and "pattern features." By consolidating these sets, we trust it is conceivable to distinguish loathe discourse: "sentiment features" permit us to extricate the extremity of the tweet, an exceptionally fundamental segment of despise discourse (given that contemptuous talks are generally negative ones). "Semantic highlights" permit us to locate any underscored articulation. "Unigram features" permit us to identify any unequivocal type of abhor discourse, while designs permit the recognizable proof of any more or verifiable types of loathe discourse.

1) Sentiment-Based Features

Fundamentally we remove features that would assist with deciding if a tweet is positive, negative or unbiased. The absolute score of negative words, and that of positive words are removed utilizing SentiStrength,6 an instrument that ascribes assessment scores to sentences just as the expressions of that it is made. The scores run from 5 to 1 for positive words, and from -1 to -5 for negative words.

2) Semantic Features

We accept that accentuation highlights, including the capitalization, the presence of inquiry and shout marks, and so forth help recognizing disdainful discourse, and they can't be basically disposed of. In our work, we utilize the accompanying highlights:

- the quantity of exclamation marks,
- the quantity of question marks,

DOI: 10.37200/IJPR/V24I8/PR280115
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

- the quantity of full stop marks,
- the quantity of all-uppercase words,
- the quantity of statements,
- the quantity of contributions,
- the quantity of laughing expressions,
- the quantity of words in the tweet.

3) Unigram Features

All unigrams that have a grammatical form (PoS) tag of a thing, action word, modifier or intensifier are removed from the preparation set and put away in three distinct records (one rundown for each class) alongside their number of events in the comparing class. We keep just words that happen in any event min^{u}_{occ} (a threshold that speaks to insignificant number of events of unigrams to be considered).

Given a word w that showed up in one of the three records (for accommodation we call it C1), we measure two proportions we allude to as ρ12 and ρ13 characterized as follows:

$$\rho 12(w)=N1(w)/N2(w) \quad (1)$$

$$\rho 13(w)=N1(w)/N3(w) \quad (2)$$

where Ni(w) is the number of appearances of the term in a class i . If the denominator of the ratio is 0, the value is set to 2.

4) Pattern Features

Pattern features are separated in a similar way we remove unigrams. The resulting vectors separated from various tweets have dissimilar dimensions; therefore, we classify a model as a vector of successive terms comprising a fixed length L where L is a factor to enhance the efficiency. If a tweet consists of more than L terms, we separate all possible patterns. If it has less term than L, it is solely removed.

iv. Algorithms Used

a) Logistic Regression

Logistic regression is basically a supervised classification algorithm. In a categorization problem, the objective variable (or input), Y, can take just discrete qualities for given arrangement of highlights (or output), X.

This algorithm becomes a classification technique only when a decision threshold is brought into the picture. The setting of the threshold value is a very important aspect of Logistic regression and is dependent on the classification problem itself. The decision for the value of the threshold value is majorly affected by the values of precision and recall. Ideally, we want both precision and recall to be 1, but this seldom is the case.

b) Naive Bayes Classification:

Naive Bayes classifiers are an accumulation of order computations reliant on Bayes' Theorem. It's definitely not a singular computation anyway a gathering of estimations where all of them share a typical rule, for instance each

DOI: 10.37200/IJPR/V24I8/PR280115
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

pair of features being arranged is self-governing of each other. Naive Bayes classifier is the least difficult and the quickest classifier. Numerous researches guarantee to have gotten best outcomes utilizing this classifier.

For a specific data, if we wish to search the name for it, we discover the probabilities of the considerable number of names, given that component and afterward select the name with greatest probability.

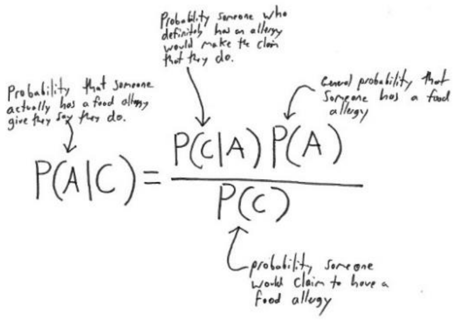

Figure 2: Formula for Naïve Bayes theorem.

c) Random Forest

Random forest is a supervised learning algorithm which is used for both classification as well as regression. Be that as it may, nonetheless, it is predominantly utilized for characterization problems. As we realize that a forest is comprised of trees and more trees implies progressively powerful forest. In the same way, random forest algorithm makes decision trees on information samples and afterward gets the forecast from every one of them lastly chooses the best arrangement by method of voting. It is an outfit technique which is better than a solitary decision tree since it decreases the over-fitting by averaging the outcome.

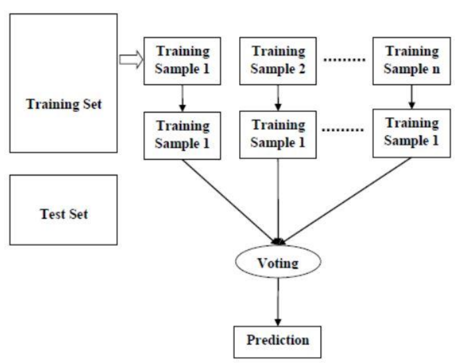

Figure 3: Prediction using Random Forest

III. RESEARCH CHALLENGES

i. Classifying Harsh Content

The order of harsh substance alludes to the class, and procedure, through which substance is distinguished as injurious and, besides, what kind of harsh material that is recognized as. This is a public and hypothetical assignment: there exists no 'right' explanation or singular criteria which should be applied. The assurance of in the case of

DOI: 10.37200/IJPR/V24I8/PR280115
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

something is injurious is likewise final to lawful explanations as these are generally moderate. Thus, utilizing the host stages' rules is frequently wrong as they are normally responsive and ambiguous. All the more by and large, the scholarly community ought to not simply acknowledge how stages outline and characterize issues as this may be affected by their business advantages.

Lucidity in sub-assignments. Distinguishing injurious substance conventionally is a significant yearning for the field. Be that as it may, it is extremely troublesome in light of the fact that oppressive material is so fluctuated. Studies that implies to direct the nonexclusive assignment of recognizing misuse is normally really tending to something much progressively explicit. This can routinely be seen from the datasets, which may contain methodical tendencies towards specific sorts and focal points of abuse. For instance, the dataset by Davidson et al. is utilized broadly for errands depicted conventionally as oppressive material location but still it is exceptionally slanted in the direction of prejudice and sexual stereotyping (Davidson et al., 2017).

Sartori's projects for the political theory on the ‘ladder of abstraction’' can be utilized to comprehend this matter (Sartori, 1970). Sartori contends about all the ideas which can be characterized, also portrayed with differing levels of deliberation. Example, 'majority rule government' can be characterized comprehensively in connection to how 'the individuals' is spoken to or barely as a lot of explicit foundations and methodology. The level of deliberation ought to be picked by thinking about the objectives and examination type – else we hazard 'swim[ing] in an ocean of observational and hypothetical untidiness.’ (Sartori, 1970, p. 1053).

Harsh substance recognition look into is as of now set apart by a lot of what Sartori names 'high' what's more, 'low' level deliberation. A few scientists use profoundly conceptual terms to depict undertakings, for example, recognition of 'misuse' or 'hailed' material. The phrases are not useful, and it is hard to understand precisely about the underlying tasks that are being tended to. For case, hailed substance might be damaging however is prone to likewise incorporate different types of non-oppressive (though restricted) content. On the opposite side, a few inquire about utilizations restricted phrases that are at an excessively 'low' measure of deliberation. Take an example, 'abhor' means a particular forceful and passionate conduct, barring different assortments of misuse, such as rejection, affront, doubt and disparaging.

Lucidity in wording. Explaining wording will assist depict the extension and objectives of study, also empower more desirable correspondence and joint effort. A portion of the principle issues are (1) specialists exercise phrases that are not that well-characterized, (2) unique ideas and phrases are utilized over the area for comparable task and (3) the phrases that are exercised are hypothetically dangerous. In particular, three parts of existing phrasing have impressive social logical constraints. The goal of the person who speaks. Oppressive substance is frequently characterized in reference to, and centers around, the speakers' expectations. Specifically, it is focal in the idea of 'abhor', which recommends a particular direction of the person who speaks. For example, Pitsilis et al. portray abhor as 'entire distributed content which is utilized to express loathe with regard to some specific gathering accompanied by the aim to embarrass its individuals' (Pitsilis et al., 2018). Somewhere else, Kumar et al. recognize 'unmistakable' from 'incognito' loathe (Ojha, Kumar, Zampieri and Malmasi, 2018). The ramifications of 'secretive' is that the people who speak are acting clandestinely to shroud their harsh aims. In any case, the aim of speakers is hard to observe utilizing socially-produced information and may not straightforwardly compare with their activities (Gillespie and Crawford, 2016; John, Margetts, Yasseri and Hale, 2015). The manner by which importance is 'encrypted' in web based settings can't be effectively found out, especially given the secrecy of numerous clients and the job of 'setting breakdown'. (Boyd and Marwick, 2010). The genuine crowd whom the person who speaks address might be not the same as the ones that they envision they are tending to (Ibid.). All things considered, little ought to be accepted about speakers' aims, and it very well may be viewed as an unsatisfactory reason for meanings of misuse.

ii. Identifying harsh content

We distinguish five phonetic challenges which increment the test of identifying oppressive substance. They have all been related with arrangement blunders in past work. Be that as it may, they are most certainly not continuously talked about and took care of methodically, and their effect is difficult to survey as they are frequently talked about subjectively instead of estimated.

Cleverness, incongruity and mockery. Evidently clever, amusing or mockery substance is frequently seen as a wellspring of order blunder. Notwithstanding, drawing on basic investigations of partiality and loathe, we suggest that such substance is as yet harsh. There are few explanation behind this. Initially, these logical gadgets have been appeared to fill in as methods for stowing away, extending and verifying authentic maltreatment. Next, people who see such substance might be unconscious of the creator and the more extensive setting, and accordingly not perceive that it’s silly – as talked about before, expectations are difficult to recognize on the web. And as far as anyone knows amusing remark which is proposed to parody misuse might be undefined from certifiable maltreatment (LaMarre, Landreville, and Beam, 2009). Third, purportedly amusing, sarcastic and funny damaging substance generally depends on a portion of partiality: the lynchpin of the expository methodology is that the crowd perceives, and maybe verifiably acknowledges, the negative tropes and thoughts related with the focused on gathering (Ma, 2014). Subsequently, while cleverness, incongruity and mockery are frequently observed as being non-abusive, we suggest that they are reconsidered.

Spelling varieties. Spelling varieties are pervasive, particularly in internet based life. Instances of spelling variety incorporate the extension of terms (for example, 'goodness' to 'ohhh') and utilization of options (for example, 'kewwl' rather than 'cool'). Spelling variety is regularly socially huge, showing articulations of character and culture. Simultaneously, a few varieties reflect semantically close indistinguishable substance (for example 'whaaaaa?' and 'whaaa?'). Spelling varieties are additionally now and again utilized adversarially to jumble and stay away from discovery (for example by utilizing surprising accentuation or on the other hand extra spaces). In most settings, it is difficult to recognize why spelling shifts.

Polysemy. This is the point at which a word with a solitary spelling has different implications. Which significance is inspired relies upon the specific situation. Luo and Magu depict how 'metaphorical' code words, for example, 'Bing' or 'Skype', are utilized to discredit specific gatherings (Joshi, Luo, and Magu, 2017). Thus, Palmer et al. portray how descriptive nominalization (for example evolving ‘Indians' to 'the Indians’) can change generally unbiased terms into disparagements.

DOI: 10.37200/IJPR/V24I8/PR280115
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

Polysemy is a specific test with damaging substance the same number of clients maintain a strategic distance from clear and obvious types of loathe (which are probably going to be consequently evacuated by stages) and rather express despise all the more inconspicuously (Daniels, 2013). Word portrayals which unequivocally consider setting are one method for conquering this issue .

Long range conditions. Many recent assessments are revolved around menial posts, for instance, Tweets (Schmidt and Wiegand, 2017). Regardless, socially made substance can cross various sentences and segments. Abuse may similarly simply be gotten through conversational components, for instance, multiuser strings (Raisi and Huang, 2016). This has been well-tended to inside investigations of cyber bullying, but on the other hand is exceptionally pertinent for the field of harsh substance recognition all the more generally. Production of progressively fluctuated datasets will address this issue, for example, utilizing information taken from Wikipedia or Reddit.

Conversation medium change. The linguistic structure, language structure, and dictionaries of language change after some time, frequently in unforeseen and lopsided ways. This is particularly legitimate with easygoing sorts of 'standard' communication medium, which duplicate in most web based spaces. One suggestion is that the presentation of frameworks prepared on more established datasets corrupts after some time as they can't represent new phonetic characteristics. Utilizing numerous transiently isolated datasets to assess frameworks will address this, just as further investigation into the effect of time on language.

iii. Representation of context

Which means is naturally questionable, contingent on the emotional standpoint of both speaker and crowd, just as the particular circumstance and control elements (Benesch, 2012). These elements have for some time been given lacking consideration in the investigation of online maltreatment, which has generally centered around simply the substance alone. This has clear restrictions.

For example, much of the time, the expression "N***a" has a practically inverse significance whenever articulated by a white contrasted with a dark individual.

Some ongoing work has begun to unequivocally represent setting by incorporating client level factors in characterization frameworks. Gambäck and Unsvåg assess a framework on three different datasets and discover that, contrasted and a standard utilizing strategic relapse with n-grams, consideration of individual-level highlights, for example, sex, informal community, profile data about data and geo-location, makes better execution (Unsvåg and Gambäck, 2018). Different investigations report comparative outcomes, utilizing both nearby and worldwide interpersonal organization highlights, perceptibly through consolidating the node2vec calculation (Labatut, Papegnies, Linarès and Dufour, 2017; Huang and Raisi, 2017). The utilization of system portrayals is bolstered by sociology inquire about which demonstrates proof of homophily on the web; all things considered, damaging clients are associated with other injurious clients (Wagemann and Caiani, 2009; Eisenberg, Tien, Porter and Cherng, 2019). We suggest that secrecy ought to likewise be unequivocally demonstrated in further project as it has bad impacts (McKenna and Amichai-cheeseburger, 2006) and is experimentally connected with clients posting misuse.

DOI: 10.37200/IJPR/V24I8/PR280115
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

The incorporation of client level highlights drives enhancements in order execution and ought to be invited as a significant advance towards more nuanced and relevantly mindful models. All things considered, we offer four admonitions. Initially, it might make transient or system investigation troublesome as the 85 grouping of clients' substance depends on these highlights, making clear danger of bewildering. Second, it might prompt new kinds of injustice and predisposition where by the substance of some of the system topologies or some hubs are bound to be distinguished as contemptuous – which may, thus, be identified with important social attributes, for example, sex or age. Third, these frameworks are to a great extent prepared on a depiction of information and don't unequivocally consider transience. It is questionable about the amount of information is desired for it to be prepared. Last, models might be lopsided by the preparation information. Wiegand et al. demonstrates about the most damaging substance in a collection of data starts from just a couple of clients by then comprising user level data dangers over fitting: the classifier just bounces on those creators' semantic properties (Ruppenhofer, Kleinbauer, and Wiegand 2019).

IV. RESULTS

After the extraction of highlights and advancement of parameters, we continue to our last examinations. The order is finished utilizing the toolbox Weka. Weka presents assortment of classifiers sorted out into bunches dependent on the kind of the calculation (e.g., choice tree-based, rule-based, and so on.).

To assess the exhibition of characterization, we utilize 4 diverse key performances indicators (KPIs) which are the level of genuine positives, the exactness, the review and the F1−score characterized as:

$$F1{-}score = 2 \times \frac{Precision \cdot Recall}{Precision + Recall}$$

Figure 4: Formula for F1 score of test’s accuracy

A. Binary Classification

In a first step, we combined the tweets of the two classes “hateful” and “offensive” under one class we refer to as “offensive” (since hateful tweets are indeed offensive and aggressive). This is to make the classification a binary classification task. In the training set, in total we have 14 000 tweets for class “offensive” and 7 000 tweets for the class “clean.” As for the test set, the number of tweets of the class “offensive” is 2,680 while that of the class “clean” is 1 340. Using these sets, we proceed with the classification.

The exhibitions per group of highlights show that the unigram includes just as the example highlights present the most noteworthy exactness with values individually equivalent to 82.1% and 70%.Semantic features doesn't have a decent characterization exactness. The equivalent goes for estimation based highlights: despite the fact that hostile language is bound to show up in negative tweets, this data alone (regardless of whether the tweet is certain or negative) isn't sufficient to decide on the substance of the tweet and the language utilized.

DOI: 10.37200/IJPR/V24I8/PR280115
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

Table 1. Binary Classification Confusion Matrix

Class	Classified as	
	Offensive	Clean
Offensive	**1174**	166
Clean	82	**588**

B. Ternary Classification

The arrangement on the test set presents an obviously lower exactness, accuracy and review. The general precision of order arrives at 79.7% with practically 10% drop in the wake of parting the class recently alluded to as "offensive" into two sub-classes (i.e., "offensive" and "hateful").

The general precision got arrives at 78.4%. Likewise, similar arrangements of highlights that performed well during the twofold order are ones that performed well during the ternary grouping, for similar reasons referenced previously.

Specifically, abhor linke unigrams are exceptionally near those hostile. As appeared in Fig. 2, words exceptionally identified with despise are nearly equivalent to those generally used to outrage individuals, disparage them or affront them (i.e., hostile discourse). That being the situation, even highlights qualified as "Unigram" present lower precision when we split the class "hostile" from the previous subsection (double characterization) into two classes which are "contemptuous" and "hostile."

Despite the fact that, performing such a correlation on designs is very testing (since designs don't demonstrate an immediate connection to a particular class), we accept that a similar sort of issue happens and the examples separated from the two classes are close and identified with each other.

Table 2. Ternary Classification Confusion Matrix of the Validation Set

Class	Classified as		
	Hateful	Offensive	Clean
Hateful	**459**	41	170
Offensive	83	**531**	56
Clean	51	7	**612**

V. CONCLUSION

Right now, proposed a new strategy to recognize abhor discourse in Twitter. Our proposed approach consequently identifies despise discourse designs and most regular unigrams and utilize these alongside wistful and semantic highlights to characterize tweets into hateful, offensive and clean. Our proposed approach arrives at a precision equivalent to 87.4% for the double arrangement of tweets into hostile and non-hostile, and an exactness equivalent to 78.4% for the ternary order of tweets into hateful, hostile and clean.

DOI: 10.37200/IJPR/V24I8/PR280115
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

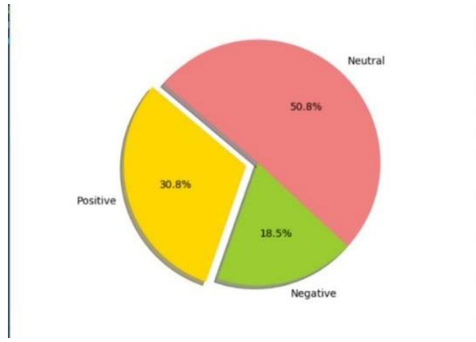

Figure 5: Pie Chart representing the occurrence of Positive, Negative and Neutral tweets

After robust data analysis, we can conclude that the majority of tweets mentioning a person are positive (30.8%), while neutral tweets (50.8%) outnumber negative ones (18.5%). However, we cannot say that the majority of people are positive, as the sum of both neutral and negative tweets (69.3%) outweighs total positive tweets (30.8%).

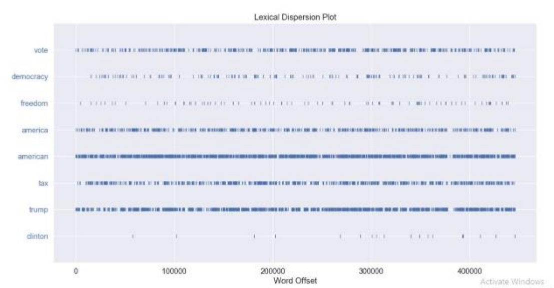

Figure 6: Lexical Dispersion Graph showing the distribution of words in the tweets

This analysis was done for the first week of April 2019, amidst great political hubbub. It is for certain that if a similar analysis was to be done at a later date, the results would vary. Sentimental analysis proves yet again to be a useful tool for gaining insight on the perception of people on various brands, personas, events etc.

REFERENCES

1. A Shervin Malmasi and Mark Dras. 2017. Native Language Identification using Stacked Generalization.
2. Shervin Malmasi, Mark Dras, and Marcos Zampieri. 2016a. Ltg at semeval-2016 task II: Complex word identification with classifier ensembles. In Proceedings of SemEval.
3. Pete Burnap and Matthew L Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. Policy & Internet 7(2):223–242.
4. Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings.
5. Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content.
6. Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection Using Natural Language Processing.
7. Huei-Po Su, Chen-Jie Huang, Hao-Tsung Chang, and Chuan-Jie Lin. 2017. Rephrasing Profanity in Chinese Text. In Proceedings of the Workshop Workshop on Abusive Language Online (ALW). Vancouver, Canada
8. Detecting Hate Speech in Social Media Shervin Malmasi, Marcos Zampieri, Harvard Medical School Boston, MA, United States, 2017.

DOI: 10.37200/IJPR/V24I8/PR280115
Received: 21 Jan 2020 | Revised: 08 Feb 2020 | Accepted: 14 Mar 2020

9. Huaishao Luo, Tianrui Li, Bing Liu, and Junbo Zhang. DOER: Dual Cross-Shared RNN for Aspect Term-Polarity Co-Extraction to appear in Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-2019), July 28 - August 2, 2019, Florence, Italy.
10. Qian Liu, Zhiqiang Gao, Bing Liu and Yuanlin Zhang. A Logic Programming Approach to Aspect Extraction in Opinion Mining. Proceedings of IEEE/WIC/ACM International Confernece on Web Intelligence (WI-2013), 2013.
11. Coulter, Ian, Gery Ryan, Lisa Kraus, Lea Xenakis, Lara Hilton, and . 2019. A Method for Deconstructing the Health Encounter in Complementary and Alternative Medicine: The Social Context. Journal of Complementary Medicine Research, 10 (2), 81-88. doi:10.5455/jcmr.20190118123212
12. Yao, L., Romero, M.J., Toque, H.A., Yang, G., Caldwell, R.B., Caldwell, R.W. The role of RhoA/Rho kinase pathway in endothelial dysfunction (2010) Journal of Cardiovascular Disease Research, 1 (4), pp. 165-170. DOI: 10.4103/0975-3583.74258