STM JOURNALS

Journal of Artificial Intelligence Research & Advances
ISSN: 2395-6720
Volume 7, Issue 3
www.stmjournals.com

Detecting Hateful Content on Social Media: A Survey

Aaniya Gouse[1,]*, Afaq Alam Khan[2]
[1]Student, Department of Information Technology, Central University of Kashmir, Jammu & Kashmir, India
[2]Assistant Professor, Department of Information Technology, Central University of Kashmir, Jammu & Kashmir, India

Abstract

Spreading hatred over social media is a growing menace. Whether in the form of discrimination, cyberbullying or trolls, malicious posts have pestered users throughout the globe. To overcome this nuisance, an automated hate text detector has almost become a necessity. Despite a few inherent limitations that hate speech detection has, there is work going on in this field, and the scope for progress remains. Besides various techniques used in the state-of-the-art, this paper extends a précis over outcomes of the latest experiments in the field. The most recent and updated approaches used in the field along with the limitations, positives of each approach and latest techniques used in the field of hate speech detection are discussed in this paper.

Keywords*: Hate Speech, Natural Language Processing, Neural Networks, Supervised Learning, Social Media.*

****Author for Correspondence*** E-mail: aaniya.gouse@gmail.com

INTRODUCTION

With the advent of social media, online interactivity has grown manifold. Owing to this, data generated on Facebook and Twitter is proliferated each minute with 350,000 tweets being sent per minute [1] and 31.25 million messages per minute being sent over Facebook, on an average [2]. Untoward incidents vis-a-vis anger and violence have proliferated along with the number of web users worldwide. Although hate speech does not have a marked definition as yet, efforts are being made to identify hate speech without bounds. Concerning social media, hate speech is a form of writing that deprecates its victim and is of serious nature. Any bigotry-driven, antagonistic or destructive statement against one or more individuals based on attributes associated with them, whether actually or seemingly constitutes hate speech. While in the US, hate speech is partly protected under the free speech provisions of the First Amendment, in countries like the UK and France, laws are put into practice that forbid hate speech, describing it as speech that singles out minority groups are encouraging barbarity or social imbalance [3]. In India, within few months, 25 incidents of mob lynching have been reported which are eventually a result of spreading of fake news. Hate speech detection should help us prevent mass hysteria, rumor-mongering over the internet, controversies, xenophobia, cyber bullying, trolling, mob violence and lynching as a cause of each of these is usually a hoax or misinterpretation of videos or other media. Also, response towards various government policies and newly launched products should be detected to be defaming, slandering or threatening. An efficient automatic hate speech detection system must report trolling, defaming, cyber bullying, etc. such that the ill effects of each of these are counter acted in time. This paper involves a brief contemplation on methods that have been incorporated to detect hate speech. Our aim is to make the new researchers acquainted with the state-of-the-art.

Out of specific challenges that we came across during the survey, one is that hate speech detection is generally limited to supervised approaches despite a shortage of labeled corpus. Despitesome hate speech detection

mechanisms being in place, there hasn't been a practical assessment of performances of present models and datasets to date. Also, there does not exist a boundary between offensive language and hateful speech. Thus, the former is often predicted as the latter. Another issue is that of considering the context of the speech. Certain hateful remarks are sarcastically formed and do not vividly appear of offensive nature. This concealed form of hate speech is a challenge in itself. Thus any automated hate speech detection system should efficentlyaddress the aforementioned challenges and present a scalable and robust system. Figure 1 depicts the general steps that are associated with hate speech detection.

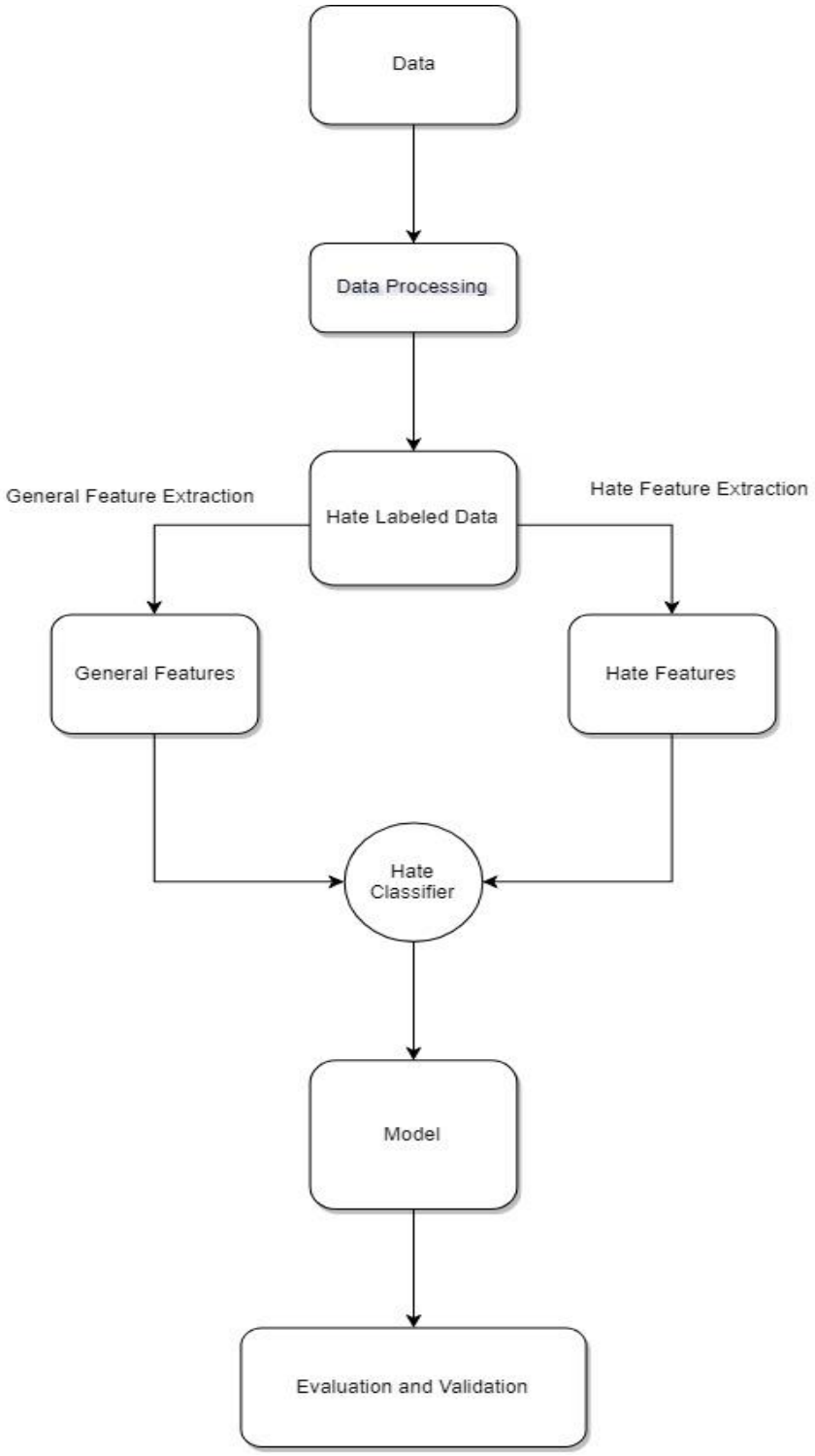

Fig. 1: *Hate Speech Detection.*

RELATED WORK

Ritesh Kumar et al. [4] grouped aggression as overt aggression and covert aggression. Bythe target of aggression, it is divided into four groups, namely, Physical Aggression, Sexual Aggression, Identity Aggression, and non-threatening aggression. They have found abuse to be an aspect of aggression as abusive statements are used to emphasize the expression of specific ideas but may or may not be aggressive. Also, as there is a line between aggression analysis and sentiment analysis, the features and characteristics for

the two are depicted differently. The tag set is so formed that it contains three tags at the top-level and ten tags at the second level. The threetop-level tags may have discursive effects and discursive roles described as an attack, defense or abet. Following certainly defined conventions for annotation, the Krippendorff's alpha for this experiment was initially 0.49(below par standard) for top-level tags. Using Crowd Flower on 1100 instances, the inter-annotator agreement rose to 72% for top level and 57% for the second level. With about 18k Tweets and 21k Facebook comments, the data is thus annotated with different levels and kinds of aggression reaching an F1 score of 0.70. It is concluded that classifying aggression even at the basic levels can be challenging.

Considering the lack of labeled corpus, a challenge, Njagi Dennis Gitariet al. [5] proposed a rule-based approach for both subjectivity analysis as well as to develop a hate speech classifier. MPQA and other sources trained subjectivity sentence detector. Using subjective sentences, semantic and subjective features are extracted. Using bootstrapping and WordNet, the lexicon is expanded with hate-related verbs and noun patterns based on target groups of religion, ethnicity, and race. Thus identified features are used to build and test the classifier. From the list of offensive sites as created by Raymond Franklin, called the Hate Directory the primary sources of the hate corpus that are used consist of a neo-Nazi website, namely, Stormfront.org and another website which contains quotes related to the Israel-Palestinian conflict. In order that the subjectivity classifier is learned both rule-based and learning-based approaches are used. Lexicon resources used are Wilson et al. and SentiWordNet. Figuratively, 3621 verbs, nouns, adverbs, and adjectives are tagged as strongly subjective. In the final experiment, 1289 words derived from theme-based patterns are included in the lexicon. The patterns with the frequency of at least twoare included on the list. A total of 103 such patterns were extracted. The annotated corpus is expanded such that various machine learning approaches such as SVM and maximum entropy are applied. These help increase the precision and recall scores.

Thomas Davidson et al. [3] have pointed out the subtle difference between hateful speech and offensive speech as a challenge in the automation of hate speech detection. Tweets about hate are collected and categorized as hate speech, offensive speech or none of these. The racist tweets are classified as hate speech while sexist tweets are classified as offensive only and as such the process is observed to exhibit a subjective bias. Also, the tweets of hateful nature but containing no hate words are not classified as hate speech. The model is trained to distinguish between each of these categories. With the seed lexicon from Hatebase.org, they expand it using tweets from 33,458 users. A random sample of 25k tweets being manually coded the inter-annotator agreement of 92% resulted in a sample of 24,802 labeled tweets. The features thus identified are used to train the classifier. The lower-cased and stemmed words are provided with TF-IDF weights. Using NLTK, unigrams, bigrams, and trigrams are formed. Various models like Naïve Bayes, Decision Trees, Logistic Regression, Linear Forests, and SVM are tested using 5-cross-fold validation only to find that Logistic Regression and SVM exhibit better performance. The final model is trained using Logistic Regression with L2 regularization. The overall precision is found to be 0.91, recall of 0.90 and an F1 score of 0.90. Misclassified hate speech is about 40% with a precision score of 0.44 and recall of 0.61. Bias is observed as tweets which are classified as hateful are only offensive. Thus, the significance of considering the context while classification is instantiated.

Zeerak Waseem et al. [6] remark that hate speech detection requires a manual effort to a great extent. Starting from a manual search of common hate words, they used Twitter to collect the corpus which is manually annotated. With men being identified as a considerable part of hate speech, the majority of users remain unidentified, thus, ruling out the possibility of using gender as a feature. Using Logistic Regression and 10-cross-fold

validation on the final model, they found that the use of character n-grams is better than that of word n-grams by about 5 F1-points. They also found that using location as a feature can be adverse on the scores. Using 2-4 grams and gender as a feature, the F1 score is found to be 73.89. Using a length of a tweet and 1-4 grams, the F1 score is found to be 73.66. Using 1-4 grams with gender and location F1 score is found to be 73.62,and with gender, location,and length it is found to be 73.47. They intend to work on improvement of location and gender classification.

Zeerak Waseem [7] states that the knowledge of an annotator about classification models affects the results of training. This paper points out that the systems that are trained on expert annotations are likely to perform better than those trained on amateur annotations. Over the best features, the model gets a high F1 score, recall, and precision but fails to classify minority classes. The dataset used suffers personal bias; false positives are, therefore, the primary cause of the error. It is suggested that weighted F1 score should be applied on the corpora. Thus preventing the misclassification of small classes.

Alena Neviarouskya et al. [8] present ways of score generation on lexicon namely SentiFul which they augmented with synonyms, antonyms, hyponyms, and composed words. Thus, they describe techniques for words implying particular sentiments. These methods can be further used to augment lexicons as it was shown to improve their performance. The seed lexicon was generated using Affect database which was then augmented by looking for new terms from synonymy relations on WordNet, new terms from antonymy relations from WordNet, by detecting already existing base words to construct new lexemes. To use this lexicon in sentiment analysis and to find the credibility of rules is set as an objective for the future.

Rafiya Jan et al. [9] presented an automated approach of sentiment analysis using the semantic similarity between two terms. Over reduced EmoLex as a resource, they used deep learning and unsupervised learning to detect semantic features from text and classified it into six basic emotion classes. The lexicon is then vectorized to finally classify the data under the labels based on the score. Augmentation of the lexicon is done using Oxford Thesaurus and WordNet. Training the Word2Vec models namely CBOW and Skip-gram model on Wikipedia dump, they derive vectors for each word. Thus, each sentence is represented by exact semantic details. Vectorization uses TF-IDF score and datasets used are Aman corpus and Affective text corpus.

The following Table 1 summarizes all the vital details that the reviewed papers contain.

APPROACHES USED

Overall, classification is all about choosing a class to which an item might most likely belong (Figure 2). In general, there are two ways, in which a classification task is performed:

(a) Multi-step classification

Multi-step classification involves learning more than one (usually two, called binary classification) classifiers. This type is used when solving problems about multiple categories. Here, we learn individual classifiers for each category termed as "one-versus-rest" [3]. For instance, the classifier learned for the whole document will decide based on the individual topics which of the polarity classifier should be used to classify unseen data [10].

(b) Standard one-step classification

Standard one-step classification involves learning a classifier for the whole of the task as opposed to a multi-step classification which divides the dataset into sub-datasets and learns a classifier for each. For instance, this will learn straight from the document, oblivious of individual topics. This way is observed to perform better than multi-step classification [10].

***Table 1**: Summary of Reviewed Papers*

Name	Approach	Features	Dataset	Results

Thomas Davidson et al. [3]	Logistic Regression, SVM, Lexicon Expansion	unigrams, bigrams, trigrams	Hatebase.org, Twitter	precision = 0.91 recall = 0.90 and F1 score = 0.90
Ritesh Kumar et al. [4]	Manual Annotation		Twitter, Facebook	F1 Score=0.70
Njagi Dennis Gitari et al [5]	Rule-based classification, Unsupervised Learning, Bootstrapping	Semantic, Semantic + Hate, Semantic + Hate+ ThemeOfExperiment	StormFront.org	Precision=73.42, Recall=68.42, F1 Score=70.83 (Best case)
Zeerak Waseemet al. [6]	Logistic Regression	unigrams, bigrams, trigrams and four grams (character n-grams with the location, gender, length of post)	Twitter	F1 Score=73.86 (gender)
ZeerakWaseem [7]	Classification	Character n-grams, Token n-grams, Token unigrams, skip grams, length, binary gender, gender probability, brown clusters, POS	Twitter	F1-Score=90.77 (Expert), 86.39(Amateur)
Alena Neviarouskyaet al. [8]	Lexicon Expansion	synonyms, antonyms, hyponyms	SentiFul, Affect database	F-Score=84.6%(positive)
RafiyaJan et al. [9]	semantic similarity, Word2Vec	Anger, Disgust, Fear, Joy, Sadness, Surprise	Aman Corpus, Affective Text Corpus	71.795%

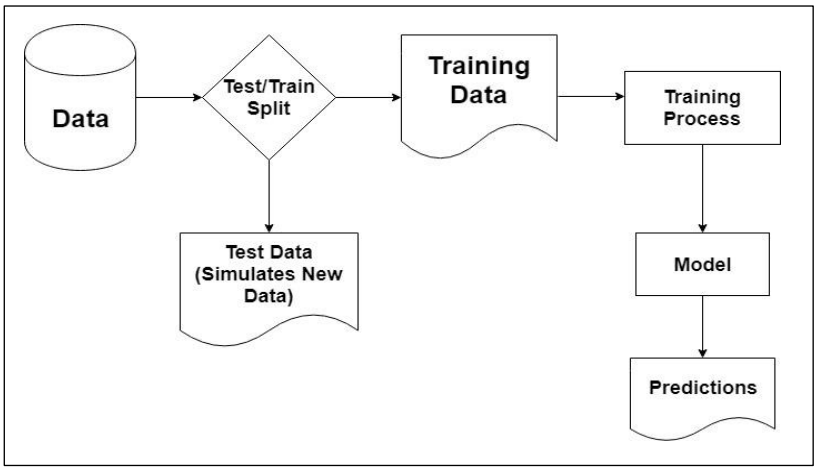

Fig. 2: *Supervised Learning Approach.*

The approaches that we came across during the survey can loosely be categorized under two groups:

(a) Machine learning based approach

Generally speaking, machine learning involves feeding some data to the computer along with some specific examples of the output that the data should produce. Based on this data and output instances, the computer learns a model. Thus given a problem the model based on some data/experience and knowledge performs in a specified way. So far we have discussed the classification mode of learning. Classification is one form of machine learning. As such, machine learning involves three different ways of learning models:

(i). Supervised learning

Essentially, supervised learning consists of two phases; the learning phase where the classifier is learned using data which is associated with their respective labels; the classification phase is that which involves determining the accuracy of the rules over test data. If found to be accurate it will be used to classify unseen data. Figure 2 illustrates the general idea behind supervised learning [11].

- *Logistic Regression:* It is used to deduct the conditional probability of an input

variable over a particular label. The outcome is a dichotomous variable, i.e. can take upto two values only say 0 or 1. Thomas Davidson et al. [5] using the one-versus-rest approach of classification for classifying the corpus into hate speech or offensive speech, realized Logistic Regression with L2 regularization performs better than Naïve Bayes, Random Forests and Decision Trees for the task.

- *Linear SVM:* A support vector machine [5] looks for the best separating hyperplane between elements of two classes. The maximum marginal hyperplane is considered to be the best one. Figure 3 illustrates the mechanism [12].

(ii). Unsupervised learning

It is used to find patterns within the data when there is no specified output. The learning algorithm is fed with just the input data out of which the similar data is grouped into clusters based on the similarities found within the data [9, 5].

- *Rule-based classification:* On satisfying a pre-condition, rule consequent is considered to be true. A rule-based classifier [5] does not involve learning and is instead based on predefined (IF-THEN) rules.
- *Neural networks:* From its roots in biology, it is used in machine learning with particular nodes being so triggered that each one produces a specific output.
- *Word2Vec:* It comprises of two models Continuous Bag-Of -Words and Skip-gram model. Both of these are based on neural networks [9] with the emphasis on semantics instead of stylistic and syntactic features. It follows a distribution hypothesis, assigning each word as a vector.

(iii). Semi-supervised learning

It uses a combination of labeled and unlabeled data to build a classifier. A classifier built for labeled data is used to make the classifier learned for unlabelled data perform better (self-training). Although their work is of unsupervised nature, Njagi Dennis Gitariet al. [1] used labeled data for evaluating their classifier.

- ***Bootstrapping:*** It involves a classifier from a small quantity of labeled data [5] later used to classify unlabeled data. This is done iteratively till the classifier teaches itself. Mainly, the accuracy of labeled data remains the focus of most research work.

(b) Lexicon based approach

Building a lexicon is based on two approaches.

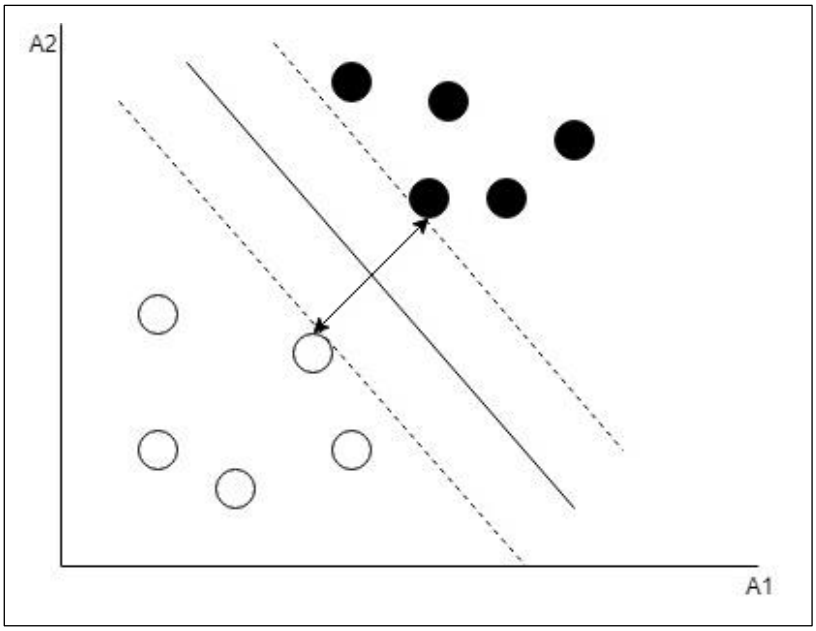

Fig. 3: *Support Vector Machine.*

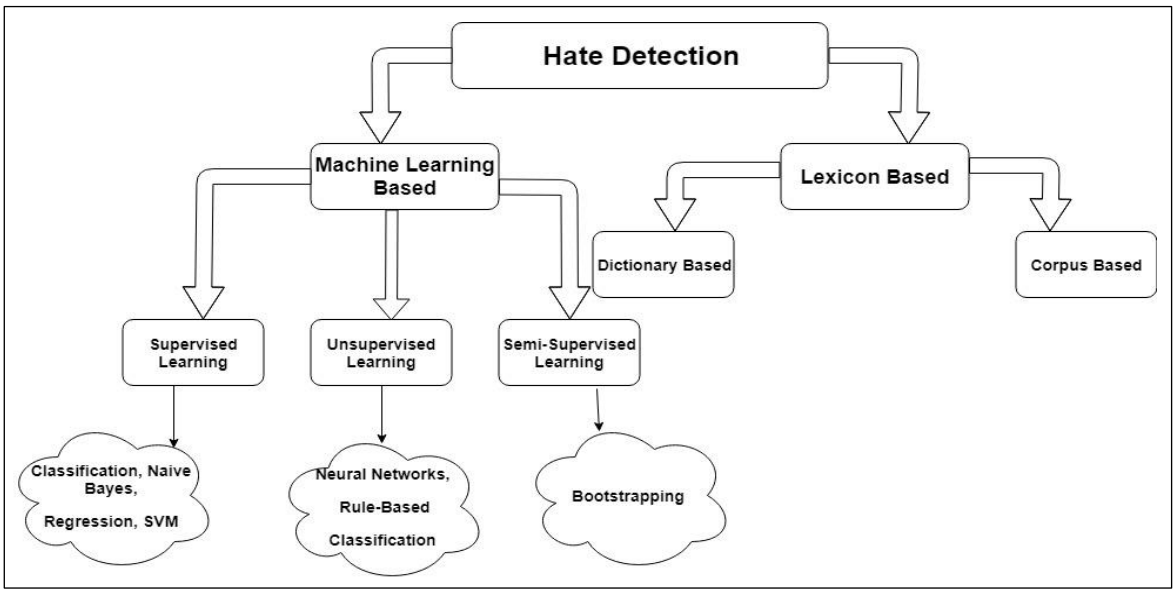

Fig. 4: *Approaches Followed.*

(i). Dictionary-based approach

This involves a static dictionary of words tagged against a polarity label and a semantic score [8]. It does not contain domain specific knowledge.

(ii). Corpus-based approach

It is a data-driven, a rule-based approach which has access to a context in addition to the sentiment labels. It has domain specific knowledge.

Figure 4 shows the broad picture of various approaches used in hate detection.

VALIDATION

Validation is performed by setting aside a part of the dataset; later used to evaluate the model. The rest of the datasetis used for training purpose. Therefore, there are three steps involved in the process of validation:

a. Setting aside of a part of sample dataset.
b. Training the model using rest of dataset.
c. Testing of model based on data set aside in (a).

Some validation techniques are [13]:

i. Leave One Out Cross-Validation (LOOCV)

Training is performed on whole of the dataset, but one small part is set aside over which it iterates. It is less biased but leads to higher variance and therefore, overfitting. Figure 5 provides an insight into LOOCV technique [13].

ii. K-Fold Cross-Validation

The dataset is divided into k number of subsets/folds over which training is performed, leaving one (k-1) part(s) for evaluating the trained model. Iteration is carried out using different subsets each time. Usually, it is performed on five subsets or 10 subsets called 5-fold cross-validation and 10-fold cross-validation [3], respectively.

EVALUATION

The most used measure of calculating accuracy is precision, recall and consequently, F-measure.

a. Precision

It is the measure of exactness of the model. It is calculated as the relevant records that we find from among the total number of records that exist in the dataset.

b. Recall

It is the measure of completeness of the model. It is calculated as the relevant records that we find from among the total number of relevant records that exist in the dataset.

There exists an inverse relation between (a) and (b).

c. F-score

It is given by the harmonic mean of Precision and Recall.

F-score=2(Precision*Recall)/Precision+Recall.

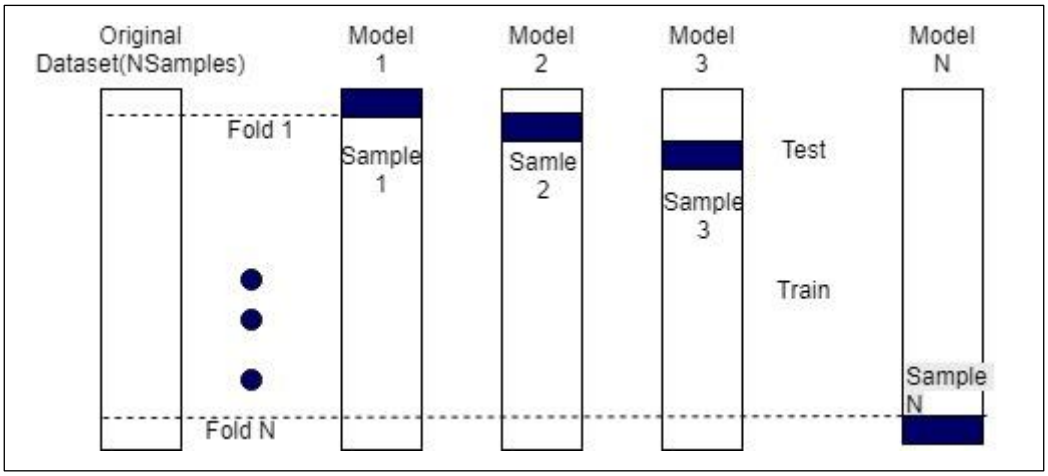

Fig. 5: *Leave One Out Validation.*

CHALLENGES

Numerous practical challenges are associated with hate speech detection which curbs the performances of existing systems. At times, hate words are so dissolved in the speech that their intended meaning remains covered up in the frequently used sense of the words. For example, use of similes and metaphors in natural language. Moreover, language that states something and implies something else is also common in human interaction. Such sarcastic languages also misclassified since it does not necessarily contain explicit hate words. As long as hate detectors remain devoid of context these issues are bound to persist.

Furthermore, hate speech does not, to date, have a strictly bound definition. It is considered in most experiments on the grounds of researchers' discretion. It requires a restricted definition which would constraint the field to one mark as this leads to the intervention of offensive language-which may/may not intend hatred, into the classified data and very often leads to inaccuracy.

CONCLUSION AND FUTURE SCOPE

Hate speech detection is mostly carried out using supervised learning algorithms. In this paper, we have summed up most of the approaches and evaluation methods that are generally employed to carry out hate speech detection. One major challenge is a loosely bound definition which often results in misclassification. It is also observed that the most efficient state-of-the-art is that which employs deep learning methods, for example, Convolutional Neural Networks. As observed the state-of-the-art methods turn out to perform below par. We intend to use subjectivity analysis to figure out the degree of anger, trolls or any other form of hatred. Since these exhibit better performance we shall implement our system using Artificial Neural Networks.

REFERENCES

1. Real Time Statistics Project (2018). *Twitter Usage Statistics*. Available from: http://www.internetlivestats.com/twitter-statistics/ [Accessed November 2018].
2. Amy Benett (2018). *7 staggering social media use by-the-minute stats.* Available from: https://www.cio.com/article/2915592/social-media/7-staggering-social-media-use-by-the-minute-stats.html [Accessed November 2018].
3. Davidson T, Warmsley D, Macy M, Weber I (2011). Automated hate speech detection and the problem of offensive language. https://arxiv.org/pdf/1703.04009.pdf. [Accessed November 2018].
4. Kumar R, Reganti AN, Bhatia A, Maheshwari T (2018). Aggression-annotated Corpus of Hindi-English Code-mixed Data.

https://arxiv.org/ftp/arxiv/papers/1803/1803.09402.pdf. [Accessed November 2018].
5. Gitari ND, Zuping Z, Damien H, Long J. A lexicon-based approach for hate speech detection. International Journal of Multimedia and Ubiquitous Engineering. 2015 Apr;10(4):215-230p.
6. Waseem Z, Hovy D. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. Proceedings of the NAACL Student Research Workshop, San Diego, California, 2016 (pp. 88-93).
7. Waseem Z. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. Proceedings of the First Workshop on NLP and Computational Social Science, Austin, Texas, 2016 (pp. 138-142).
8. Neviarouskaya A, Prendinger H, Ishizuka M. SentiFul: A lexicon for sentiment analysis. IEEE Transactions on Affective Computing. 2011 Jan;2(1):22-36p.
9. Jan R, Khan AA. Emotion Mining Using Semantic Similarity. International Journal of Synthetic Emotions (IJSE). 2018 Jul 1;9(2):1-22p.
10. Park JH, Fung P. (2017). One-step and two-step classification for abusive language detection on twitter. https://arxiv.org/ftp/arxiv/papers/1706/1706.01206.pdf. [Accessed November 2018].
11. Valletta JJ, Torney C, Kings M, Thornton A, Madden J. Applications of machine learning in animal behaviour studies. Animal Behaviour. 2017 Feb 1;124:203-220p.
12. Jiawei Han, MichelineKamber, Jian Pei. Data Mining Concepts and Techniques. 3rd ed. Massachusetts: Morgan Kaufmann Publishers; 2011.
13. Coursera. Applied Machine Learning in Python. Available from: https://www.coursera.org/lecture/python-machine-learning/cross-validation-Vm0Ie [Accessed 12 November 2018].

Cite this Article
Aaniya Gouse, Afaq Alam Khan. Detecting Hateful Content on Social Media: A Survey. *Journal of Artificial Intelligence Research & Advances*. 2020; 7(3): 1–9p.