# Hate Speech Detection in Twitter-A Survey

*Neha Naaz Y. Malik [*1], Krishnakant P. Adhiya [2]*

*[1] Department of Computer Engineering, SSBT's COET Bambhori*
*[2] Department of Computer Engineering, SSBT's COET Bambhori*
*[1]maliknehanaaz@gmail.com, [2]kpadhiya@yahoo.com*

***Abstract***

*Online social networks (OSN) and microblogging websites are attracting internet users more than any other kind of website. Services such those offered by Twitter, Facebook and Instagram are more and more widespread among individuals from different backgrounds, cultures and interests. The communication amid people from all kinds of cultural and psychological backgrounds are very much straight through as the social networks and microblogging websites are growing very rapidly, leading to more and more cyber conflicts between these people. Hate speech can be defined as usage of unnecessary aggression, violent or offensive language, which are pointing on particular group of individuals sharing common characteristics, which may include their believes , gender, religion or ethnic group . In this paper, we have given the survey of various techniques for Hate speech detection.*

***Keywords****: Hate speech, social network, sentiment analysis, machine learning.*

## I. Introduction:

Whenever clients draw in on the web, regardless of whether on message board gatherings, online chats, blogs or web based life, there is always constant probability that these activities can end up hurting individual's sentiments due to different kind of remarks made by users. Words and sentences such as "Fuck off idiots", "Go! Get life you pervert", "These people are insane", "Impotent", "Assholes" etc. are sadly normal on the web and can profoundly affect user or society's courtesy. Hate Speech is typically plot as "Criticizing individual or a group based on some common things it may be race, complexion, civilization, gender, sexual orientation, nationality, religion, or other trademark." Hate speech is unfortunately basic event on the Internet and at times wind up in extreme dangers to people.

Strong association between hate speech and actual hate crimes raises the significance of detecting and tempering hate speeches. Recent cases bring to light the effect of using cruel utterance on social media and major firms. As an example, in 2013, Facebook experienced harsh criticism for facilitating pages which were contemptuous against ladies such as viciously assaulting a companion only for snickers and kicking girlfriend since she won't make a sandwich. Facebook isn't the sole company that contends with these issues; any company that allows users to post their stuffs will have to face issues. This highlights the significant impact a community or a company have to undergo because of hateful language.

### 1.1 Data Preprocessing:

Data Preprocessing is a technique of converting raw data into a clean data set. For detecting hate speech steps for data preprocessing include:

- The Removal of URLs (starting either with \http://" or \https ://") and unrelated expressions (words written in languages that is not supported by ANSI coding).

- Replacing word with contraction: Text normalization is necessary in pre-processing to rectify the errors in text or words. Contractions like didn't, couldn't are common in tweets which are replaced with original form.
- Elongation Replacer: Removing repeated characters to end up with a regular meaningful word is an important step in text normalization. e.g. haateee
- Tokenization: Emoticons, slangs and acronyms present in the tweets are strong indicators of emotion or sentiments. Tokenizer captures all these things.
- Lemmatization: It is used to get a valid meaningful root word.
- Part-of-Speech (PoS) Tagging: In this a word is marked in a text according to the part of speech it belongs, based on both its definition and its context.
- Negation replacer first identifies the occurrence of negated word like 'not' and then replacer will find an antonym for the next word and replace it if an unambiguous antonym is present for the word. A word with more than one antonym won't get replaced.

**1.2 Feature Extraction:**

Following are some features that can be extracted for detection of hate speech.

- Sentiment-Based Features: It is used for detecting polarity of a tweet.
- Semantic Features: Punctuation features, including the capitalization, the existence of question and exclamation marks, etc. helps detecting hateful speech.
- Unigram Features: Unigram features are simply unigrams collected from the training set and are used each as an independent feature.
- Pattern Features: Words are classified by its polarity along with pos tag.
- Knowledge-Based Features: This feature makes use of world knowledge to categorize tweet as hateful which would otherwise be considered as clean.
- Linguistic Features: Knowing dependency relationship between offensive terms and hate target is beneficial to classify the tweet in hate category.
- Lexical Resources: Presence of specific negative words (such as slurs, insults, etc.), are considered as a feature. A popular source for such word lists is the internet.
- Meta-Information: Meta-information is also advantageous in detection of hate speech. Background information about the user of a post also contributes in identifying, as a user who writes hate speech messages is more likely to do so again in future than the one who does not.

## II. Literature Survey

Kwok and Wang [1] targeted the detection of hateful tweets against black people. Training dataset is built containing sample tweets which have already been classified and contain concurrent features. Racist twitter accounts were used to select racist tweets and vocabulary is constructed from the processed tweets using unigram features in training dataset.

Gitari et al. [2] precedes the approach in three steps. The first step involves subjectivity detection, and is meant to segregate sentences having subjective expressions from sentences showing objective sentiments. Second step used rule-based method for building lexicon of hateful words considering subjective features identified in step 1 and semantic features learned directly from the corpus. Finally, classifier is trained with the help of features created from lexicon and used for testing hate speech in a document.

Burnap and Williams [3] used typed dependencies (i.e., the relation between words) along with bag of words (BoW) features to distinguish hate speech utterances from clean speech ones. Supervised machine learning method was trained and tested using 10-fold cross-validation approach.

S. Sood and E. Churchill [4], observed that some black listed words might not be offensive in proper context, as most basic approaches make use of predefined black-lists only, so they used edit distance metric as well to improve performance of profanity detection. This resulted in identifying un-normalized terms such as ass or sh1t. Crowdsourcing was used first time by them to annotate abusive language.

Waseem and Hovy [5] used non-linguistic features including gender and ethnicity of author. For annotating a publicly available tweet corpus of more than 16k tweets, they used a list of criteria found in critical race theory and analyzed the impact of using character n-grams along with various extra-linguistic features for hate speech detection. Performed grid search over all possible feature set combinations to pick the most suitable features, found that if character n-grams are used it outperforms word n-grams by at least 5 F1-points.

Nobata et al. [6] employed a supervised classification method which used NLP features to measure different aspects of the user comment. They experimented with different syntactic features along with different types of embeddings features, which resulted in improved performance when combined with the standard NLP features.

Y. Chen, Y. Zhou [7], they used lexical and parser features together to detect offensive language in YouTube comments to protect teenagers. And they were the first one to use this combination of features. Parents or teachers could adjust the tool using a threshold so that contents will be filtered out before appearing online. Support Vector Machines (SVMs) of supervised classification is used with features including manually developed regular expressions, n-grams, blacklists along with dependency parse features.

N. Djuric, R. Morris [8] proposes a method which can be divided in two steps for hate speech detection. First step include the use of paragraph2vec for joint modeling of comments and words. Then, binary classifier is used which is trained with the help of embeddings to distinguish between hateful and non-hateful comments.

Warner and Hirschberg [9] used support vector machine for detecting hate speeches. Occurrence of words in a 10-word window, brown clusters and word n-grams, were used to train a model. And found that their model produces unigrams as most indicative features by resulting in F1 score = 63, which is same as the F1 score that can be obtained by word n-grams.

Pang and Lee in [10] used subjectivity detector which is meant to segregate sentences having subjective expressions from sentences showing objective sentiments. Then they used traditional bag-of-words features in conjunction with inter-sentence level contextual information using minimum cuts formulation. Their model showed considerable improvement over a baseline word vector classifier.

Dinakar et al. [11] presents an approach which focused on anti-LGBT hate speech with the help of world knowledge. ConceptNet is used for this purpose. Assertions were formed by encoding concepts that are connected by relations. Social media network Formspring was used to augment ConceptNet with a set of stereotypes which were extracted manually, and it is named as BullySpace. This knowledge base is used to compare similarity between concepts of common knowledge and concepts expressed in user comments. Then similarity between

four canonical concepts and extracted concepts is calculated. The resulting similarity score indicate if the message is hate speech or not. This approach is only applicable for a very restrained subtype of hate speech which is anti-LGBT bullying.

T. Davidson, D. Warmsley, M. Macy and I. Weber [12] used logistic regression with L1 regularization for data preprocessing. And as a final model used logistic regression with L2 regularization to examine the predicted probabilities of class membership, which is trained using entire dataset to classify each tweet.

Vigna et al. [13], for Italian language they were the first to design and develop hate speech classifier. For sentiment analysis they compared two different state-of-the-art learning algorithms. Sentiment polarity and word embedding lexicons were used to improve overall accuracy of a system.

Xiang et al. [14] propose an approach that exploits the lexical emplacement of profane language via statistical topic modeling techniques and in a single machine learning framework detects offensive tweets using significant topical features as well as the reliable lexicon feature.

**Table 1. Survey Table for Hate Speech Detection**

| Year | Authors | Approach used | Features extracted | Strength and Limitation |
|---|---|---|---|---|
| 2017 | Vigna et al. [13] | Support Vector Machines and Long Short Term Memory | Lexical features Syntactic Features Word embeddings | The outcome shows that this Hate Speech corpus, allows building automatic hate speech classifier which is able to achieve accuracy as obtained from subjectivity and polarity classification for Italian language. |
| 2016 | Waseem and Hovy [5] | logistic regression classifier | Demographic, geographic and lexical distribution | To spot racist and sexist slurs a list of criteria is prepared by using critical race theory which could collect huge amount of data and address the problem of a small, but highly abundant number of hateful users.<br>Non linguistic information is often unavailable or unreliable on social media. |
| 2015 | Burnap and Williams [3] | SVM, BLR, RFDT | BoW and syntactic features, | The classification results in very high level of performance at reducing false positives and produced promising results with respect to false negatives. |
| 2015 | Gitari et al. [2] | rule-based classifier | semantic features and grammatical patterns features | Performance is improved by using semantic, hate and theme-based features together. Precision and recall are improved with the use of |

| | | | | |
|---|---|---|---|---|
| | | | | subjective sentences. |
| 2013 | Kwok and Wang [1] | Naïve Bayes classifier | unigram features | Built dictionary of unigrams (as it contains terms related to black people only) cannot be reused to detect hate speech towards other groups with same efficiency. |
| 2012 | Dinakar et al. [11] | supervised machine learning classification | bag-of-words | Constructed a common sense knowledge base named as that encodes particular knowledge about bullying situations. This approach only works for LGBT bullying. |
| 2012 | Xiang et al. [14] | J48 decision tree learning, Support Vector Machines (SVM), logistic regression (LR) and random forest (RF). | Lexicon feature | The experiment results suggest that proposed approach is able to detect up to 5.4% more profane patterns without sacrificing the FP, which is a statistically significant improvement and is of great practical importance. |
| 2012 | Warner and Hirschberg [9] | support vector machine classifier | Unigram features | System resulted in low recall showing there are larger linguistic patterns that their shallow parses could not detect. |
| 2012 | Y. Chen, Y. Zhou [7] | Naïve Bayes (NB) and SVM | Lexical, style ,structure and context-specific features | To identify offensive content and offensive users in social media, Lexical Syntactic Feature (LSF) based framework is used. |
| 2004 | Pang and Lee [10] | Naive Bayes and SVM | bag-of-words features | Performance is significantly improved as compared to baseline word vector classifier. |

## III. Conclusion

As social media allows users to share their contents, the amount of data is growing day by day. It is necessary to use accurate, automated methods to detect hatred in online contents. User can abandon online community if problem is not addressed. In this paper, we presented a survey on techniques for detecting hate speech. Also some existing systems are discussed above which has their own advantages, features and limitations.

## References


[1] Kwok and Y.Wang, “Locate the hate: Detecting tweets against blacks," in Proc. AAAI, pp. 1621_1622. Jul. 2013

[2] N. D. Gitari, Z. Zuping, H. Damien, and J. Long, “A lexicon-based approach for hate speech detection," Int. J. Multimedia Ubiquitous Eng. vol. 10, no. 4, pp. 215-230, Apr. 2015.

[3] P. Burnap and M. L. Williams, “Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making," in Policy Internet, vol. 7, no. 2, pp. 223-242, Jun. 2015.

[4] J. A. S. Sood and E. Churchill, “Profanity use in online communities," Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1481-1490, 2012.

[5] Z. Waseem and D. Hovy, “Hateful symbols or hateful people? Predictive features for hate speech detection on twitter." RW@HLT-NAACL, pp. 88-93.

[6] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, “Abusive language detection in online user content," in Proc. WWW, pp. 145-15 , Apr. 2016.

[7] S. Z. Y. Chen, Y. Zhou and H. Xu., “Detecting offensive language in social media to protect adolescent online safety." in Privacy, Security, Risk and Trust (PASSAT), pp. 71-80 2012.

[8] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, “Hate speech detection with comment embeddings," in *Proc.* WWW Companion, pp. 29-30 , May 2015.

[9] W. Warner and J. Hirschberg., “Detecting hate speech on the world wide web," Proceedings of the Second Workshop on Language in Social Media, Association for Computational Linguistics., pp. 19-26, 2012.

[10] B. Pang and L. Lee., “A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," Association of computational linguistics, 2004.

[11] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard “Common sense reasoning for detection, prevention, and mitigation of cyberbullying.” ACM Trans. Interact. Intell. Syst, September 2012.

[12] T. Davidson, D. Warmsley, M. Macy, and I. Weber, “Automated hate speech detection and the problem of offensive language," in Proc. ICWSM, pp. 1-4, May 2017.

[13] Fabio Del Vigna, Andrea Cimino, Felice Dell’Orletta, Marinella Petrocchi, and Maurizio Tesconi. “Hate me, hate m not: Hate speech detection on facebook.” in Proceedings of the First Italian Conference on Cybersecurity, pages 86–95, 2017.

[14] Guang Xiang et al. “Detecting offensive tweets via topical feature discovery over a large scale Twitter corpus.” in Information and Knowledge Management, pages 1980–1984, 2012.