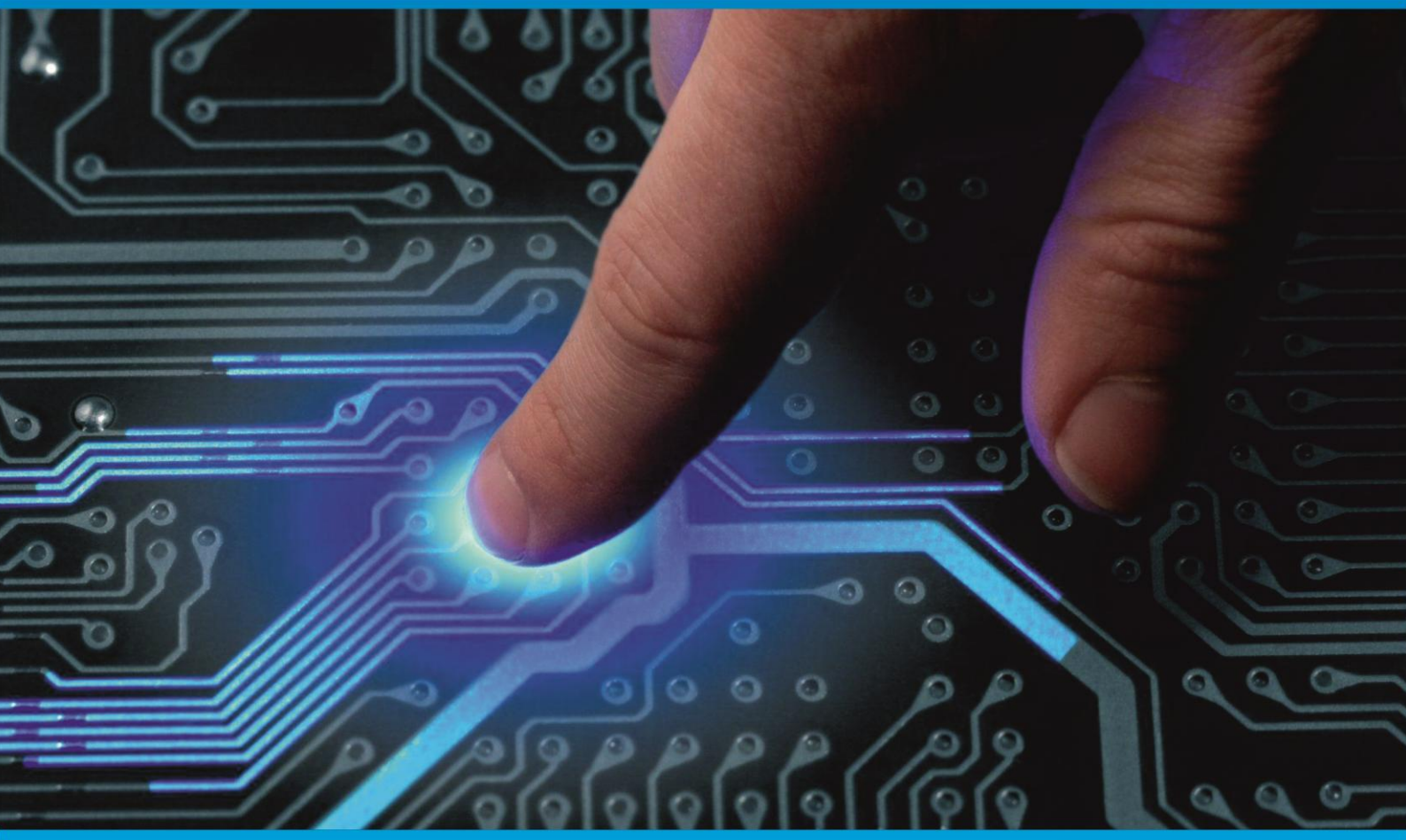
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 8, Issue 11, November 2020

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.488

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 11, November 2020 ||

An Enhanced Method of Offensive Text Detection

Shalinee Kushwaha, Dr. Nitesh Dubey

PG Student, Dept. of CSE., GNCSGOI Jabalpur, RGPV University, Bhopal, India

Assistant Professor, Dept. of CSE., GNCSGOI Jabalpur, RGPV University, Bhopal, India

ABSTRACT: With technology developing in a incredible speed, more and more people are no longer limited by time and space and get themselves involved with the massive internet. Users usually have unrestricted access to huge amount of online content without paying much effort, which brings them unimaginable benefits along with vast amount of online bullying. Due to Internet's non-restrictive nature and certain countries' legal protection of free speech also including hate speeches, some users take advantage of these manners to spread hatred and offensive language online, bringing disgusting social media experience toward other users. To mitigate these harmful influences, some social media platforms such as Facebook and Twitter have announced that they would try to address these problems.

In this thesis, we present a simple though robust machine learning method to detect targets among offensive speeches. Our approach outperforms many machine learning methods, including the official bidirectional Long Short Term Memory model, while requires significantly less time as well as resources for training and produces much more explainable decisions.

KEYWORDS: Sentiment analysis, Twitter, Adjective analysis, Linear SVM, Logistic Regression, Hate Words, Machine Learning.

I. INTRODUCTION

Automatic detection of hate speech has become an increasingly relevant research topic in the past few years [1]. The worldwide adoption of online social networks has created an explosion in the volume of text-based social exchanges. Social media communications can strongly influence public opinion and some social platforms are said to have enough social capital to influence the outcome of democratic processes [2]. Therefore, correctly assessing hate speech and other forms of online harassment has become a pressing need, to guarantee non-discriminatory access to digital forums, among other things [3]. Large social media providers, such as Facebook and Twitter have mechanisms for users to report hate speech. However, this approach requires efficient automatization techniques for the evaluation of such content, which does not appear to be simple: user accounts that constantly post potentially dangerous hateful expressions have incorrectly been deemed as harmless, and blatantly offensive content can go unreported for long periods of time. Given the enormous volume of content posted daily in these platforms, human editorial approaches have become unfeasible. Hence, the incorrect assessment of toxic content can be most likely attributed to the lack of reliable mechanisms for its automatic detection. Twitter, for example, has publicly declared its commitment to "serve healthy conversations" and "to help increase the collective health, openness, and civility of public conversation, and to hold ourselves publicly accountable towards progress."

1. Among other things, Twitter has even announced funding initiatives for academic research on this topic.

2 Despite the apparent difficulty of the hate speech detection problem evidenced by social-media providers, current state-of-the-art approaches reported in the literature show near-perfect performance. Within-dataset experiments on labeled hate speech datasets using supervised learning achieve F1 scores above 93% [4]. Nevertheless, there are only a few studies towards determining how generalizable the resulting models are, beyond the data collection upon which they were built on, nor on the factors that may affect this property [5]. Furthermore, recent literature that surveys current work also views the state-of-the-art under a more conservative and cautious light [5].

In this work, we take a close look at the experimental methodology utilized for achieving the results described by the state-of-the-art methods. We focus on two methods reporting the best results for hate speech detection over Twitter data: the work by Badjatiya et al. [4] (93% F1 – WWW 2017), and by Agrawal and Awekar [6] (94% F1 micro and macro-average F1 – ECIR 2018). At first, our intention was to replicate these findings to then measure how these models would perform on similar yet different datasets. However, a closer look at the papers and the code provided by the authors for replicating experiments, revealed details in their implementation which can produce data overfitting. In both cases there were very subtle issues that are not directly apparent from the description of the methods or from the companion code. For the case of the work by Badjatiya et al. [4], the issue is produced in the way that the authors compute features from the input data. In the work by Agrawal and Awekar [6] the issue is produced by how the authors perform the oversampling of the minority classes. To study the effects of the aforementioned methodological issues that we observed in prior work, we replicated these methods exactly as presented by the authors. This was done to

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 11, November 2020 ||

ensure we could obtain their reported performance using their code and the same data. Next, we made corrections to avoid data-overfitting and re-evaluated the generalization error of such approaches.
In summary, our work shows that although state-of the-art methods report impressive performances [6]; hate speech detection is far from solved in mono and cross lingual scenarios. We provide an explanation for this by exposing some methodological issues, but also by showing the impact of some inherent biases in the datasets that are publicly available and widely used [7]. In light of our findings we believe that it is important to pay careful attention to experimental evaluation and how predictive models generalize.

1.1 Hate Speech on Social Media:

Hate speech will act as an obstacle to these goals. The impact of hate speech is not same in all instances, depends on the person involved, content, location, and circumstances. This indicates that who, what, where and a circumstance determines the impact of a hate speech and its control. Hate speech may harm the victims directly or indirectly. In direct hate speech, the victims are injured immediately by the contents of hate speech. In an indirect hate speech, the harm may be immediate or delayed; the delayed harm is perpetrated by the agents, not by an original actor. For instance, the hate speech on racism in public meetings might motivate other racists to initiate harassment, intimidation, violence and so on (Seglow, 2016).
Figure 1.1 shows the role of online social networks for destructive activities such as hate speech, hate crime, extremism, and terrorism.

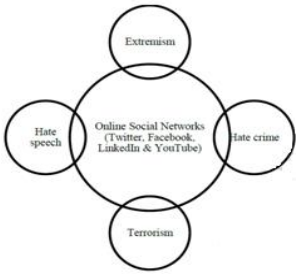


Figure 1.1: Role of OSN for destructive activities.

Hate speech is made spreadable by posting a message, reposting a message and responding to a message on social networks. Hate crime is a hate-motivated physical attack and social networks are used for planning and executing the attack related activities. Extremists and terrorists use social networks for contacting and recruiting like-minded persons, spreading propaganda, planning and executing the attacks. Hate speech, immediately after the event (influence stage) will flow heavily on social networks, after few days (intervention stage) will get reduced, after some more days (response stage) reduces to zero level and after a long time once again it may appear. This indicates that after a particular event people will be more excited and gradually will get a normal state or behavior. The rebirth stage is shown with a dashed line to indicate as an optional stage. Based on the type and impact of an event, the hate speech may or may not appear once again after a long time.

II. RELATED WORK

2.1 Definition of Hate Speech

Hate speech generally targets ignorant groups to exhibit an opposing behavior on them. The superiors will forget that the ignorant group will also have an equal right while making hatred statements. Hate speech is more destructive and dangerous when it targets traditional symbol, event or an activity. The messages exchanged on individuals related to nation, race, ethnicity, religion, sexual orientation, occupation, gender or disability have a more impact than the individuals personal information. [8] Has defined hate speech "as bias motivated, hostile, malicious speech aimed at a person or a group of people because of some of their actual or perceived innate characteristics". The European Court of Human Rights, adopted a definition on hate speech as "all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility towards minorities, migrants and people of immigrant origin" (Council of Europe, 1997. With this concept, we assume that "hate speech is any speech, which attacks an individual or a group with an intention to hurt or disrespect based on identity of a person". Once the hate speech is expressed, hurting or disrespecting depends on the perception of the victim. For some, it may or may not affect. Generally, an effect of hate speech depends on the originator, content and the targeted one.
If a hate speech does not incite to discriminate (do not hurt the targeted one), then, there arises a question that whether this kind of speech is hatred or not? Here it is accepted as hate speech because of the intention and content. For clarity consider a legal framework, in which an attempt to murder is treated as a crime, accused will be penalized and the

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 11, November 2020 ||

victim will be provided more protection. Here purpose and action performed by the murderer are counted. Similar ideology is applicable in the context of hate speech.

As a part of the legal frameworks, some of the commonly acceptable activities related to expressions like free speech and hate speech by national and international bodies are discussed. The legal frameworks contain set of rules to permit or prohibit activities or ideas based on their nature. The legal information on hate speech can be found by accessing international human rights law with internationally accepted declarations and conventions supporting fundamental rights to every human being. Article 19 from Universal Declaration of Human Rights (UDHR) states that **"**Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any networks and regardless of frontiers". The whole universe is agreed upon the freedom of expression. To make effective and appropriate use of freedom of speech, article 29(2) of the UDHR states that, "In the exercise of his rights and freedoms, everyone shall be subject only to such limitations as are determined by law solely for the purpose of securing due recognition and respect for the rights and freedoms of others and of meeting the just requirements of morality, public order and the general welfare in a democratic society." It opposes the use of text, content, theory, and practice of free speech as a liberty of an individual in the modern societies. Similarly, other international bodies stated their views on free speech rights and/or hate speech restrictions in the form of articles.

2.2 Uses of Machine Learning in Hate Speech Detection

The field of hate speech automatic detection and classification has evolved rapidly in the past years. Interest has increased as social media and social platforms have grown in terms of influence and user adoption. Similarly to the field of sentiment analysis, automatic hate speech identification research has stemmed from two types of approaches: those based on the use of lexicons and those based on machine learning. In this literature overview we focus on machine learning approaches for detecting hate speech in social media textual content. Most work focuses on hate speech detection for Twitter messages, also known as tweets, which are short text messages. Thus, we also focus our review on works that use this type of data. Several previous works have tried a diverse range of classic machine-learning strategies. These usually have an initial feature-extraction phase, such as computing Term-Frequency Inverse-Document-Frequency scores or Bag-of-Words vectors, but also combining it with Meta information such as information from the user account and information about the network structure (followers, replies, etc.) [9]. These features are then used as input for methods such as Logistic Regression, SVM, or Random Forest classifiers. More recently, Deep Learning methods have attracted interest to hate speech detection [10]. As opposed to more traditional methods, Deep-Learning methods are able to automatically learn representations of the input data that can be used as features to classify it.

Although most of the proposed models have been developed in monolingual scenarios, there are some multilingual approaches in the literature. Saha et al. [11] proposed a language-agnostic model for hate speech detection using multilingual embedding representations (BERT and LASER) and a Gradient-Boosted Decision Tree classifier. This model was tested separately using datasets in Hindi, English, and German languages, reporting a 78% for the Hindi dataset. For the SemEval 2019, multilingual approaches were proposed for identifying hate speech in tweets in two different tasks for English and Spanish languages, separately. In particular, Bojkovsky and Pikuliak [12] used LSMT and CNN architectures jointly with multilingual embedding representations. They trained the models with a concatenation of English and Spanish training sets to classify both testing sets, but the results did not improve compared to their monolingual counter-parts. Agrawal and Awekar [13] test the performance of models trained on tweets classifying on Wikipedia data and Form spring data. The authors show that transfer learning from Twitter to the two other domains performs poorly achieving less than 10% F1. In a similar study, Dadvar and Eckert [14] perform transfer learning from Twitter to a dataset of YouTube comments showing a performance of 15% F1. Grondahl et al. [48] present a comprehensive study reproducing several state-of-the-art models. Especially important for us is the experiment transferring Badjatiya et al.'s model trained on the Waseem and Hovy's dataset to two other similarly labeled tweet datasets. Even in this case the performance drops significantly, obtaining 33% and 47% F1 in those sets. This is a 40+% drop from the 93% F1 reported by Badjatiya et al. From these results, Grondahl et al. [15] draw as a conclusion that model architecture is less important than the type of data and labeling criteria being used. In this paper our results are coherent with those of Grondahl et al. [15]. However, we take our research a step further by investigating why this issue occurs.

2.3 Categories of Hate Speech

Hate speech does not target based on only single identity. It can target on the basis of gender, religion, race, and disability [16]. In the following subsections, a review of hate speech based on gender, religion, race, and disability is made.

2.3.1. Gendered hate speech

This is an expression, which is made on the grounds of gender or sex. The victims of this kind of hate speech are generally women and girls. There is an intended violence on women and girls in the world due to their gender identity.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 11, November 2020 ||

This is known as sexist hate speech and is a kind of social shaming which intends to disrespect women, introduce fear and insecurity among women in the society. Easy availability of the Internet, the rapid growth of information and communications technologies and the common use of social networks made depicting violence against women and girls much simple. These advancements are being used as tools to harm women and girls. Online violence against women and girls is considered as a global problem. Social networks are the primary medium for an online harassment on the basis of gender. This kind of harassment with women affects personal lives and professional careers of women. Both women and Muslims are targeted by online hate than any other gender and community. For the academician who faces societal inequalities such as women or a person belonging to Muslim community, the internet may be unsafe space. An abuse and harassment of the women and girls in the society might be the one of the reason for a female to move towards terrorist organizations.

Hate crimes are increased by legal inequalities because they lead to biasing and violence. Violence can be reduced with legal equalities. [17] Highlighted that there is a need to have analytical research for providing insights to empower victims, to discourage perpetrators and to increase awareness among the public. Barlow suggested that the social networks companies, like Twitter, should take corrective measures to counter online abuse against women and Muslims. [18] Identified that women are recruited by terrorist organizations mainly to meet sexual requirements of the men. Based on the identified relationship among the predictors of traditional bullying and cyberbullying, suggested that educational programs can be used as a tool to counter abuses of both bullying and cyberbullying. Factors involved such as personality, contextual and roles are closely related to both the acts.

Beckman et al. [19] determined the role of youngsters with gender differences engaged in traditional bullying and cyberbullying using data samples of size 2989 from school students of Sweden to control cyberbullying Bastiaensens et al. (2014) examined the effect of contextual factors on bystander's behavioral intentions towards helping the victim or reinforcing the bully during the harassment using Facebook with the data collected from 453 secondary school students of Flemish.

After analyzing the attitude towards gender, a statement such as women are dedicated caretakers and mothers and men are facility providers are made by Ridgeway (2011). Similarly, [19] identified the nature of women and men towards contact establishment with others in the society. Levy and Levy after analyzing the effects of 3 policies on a partnership of same-sex, non-discriminated employment and laws of hate crime with annual data from 2000-2012, shown that hate crimes are affected by public policies related to sexual orientation. Hardaker and McGlashan (2015) investigated the sustained period of abuse and harassment towards a feminist campaigner and journalist, Caroline Criado-Perez via her Twitter account using an interdisciplinary approach with quantitative and qualitative analysis. Jane (2016) examined the responses of feminist to increasing problems of online hate with a focus on female gamers and the responses of Australian gamer Alanah Pearce with alert messages to their mothers against sexual violence threats from young male Internet users.

3.2. Religious hate speech

This is a type of hatred expression against religions such as Islam, Hindu, and Christian. As the religion contains the group of people, the hate speech against this is more harmful than against an individual. Muslims are demonized and vilified online with negative attitudes, stereotypes, discrimination, physical attacks and harassment with an intention of creating violence. Anti-Muslim abuse is increasing online, so it is required to address Islamophobia issue on social networks. An analysis of online communities is possible by observing their activities such as information they post, share and like [20]. Muslims are being used as a model to depict homogeneous out-group which is involved in conflict, violence and extremism. The internet acts as an amplifier to reflect and reinforce available discourses into networks for stronger polarized effects.

3.3. Racist hate speech

An expression towards the appearance of a person or group is known as racist hate speech. Usually, this kind of speech takes place at international level. The frequency of occurrence and impact of this speech depends on the intention and perception of the government of a particular nation and varies from one leadership to another leadership. Tatum has argued that, "racism as a system involving cultural messages and institutional policies and practices as well as the beliefs and actions of individuals" [58]. Wodak and Reisigl assumed that "racism is both an ideology of a syncretic kind and a discriminatory social practice that could be institutionalized and backed by the hegemonic social groups". This indicates that, in an environment or a system, people of one group exhibit their power against other group/individual based on physical appearance such as skin color.

3.4. Hate speech on disability

The incitement made against the physical and mental conditions of a person is referred as hate speech on disability. Disability is considered as a social category like race and gender rather than perceived as an isolated entity of medical field. Disability means any health problem of an individual which limits to do some of the life activities. With the presence of advanced medical diagnosis and treatment, the people survive longer with the help of supporting tools but results in disability. Disability can be a part of any person, at any time of the life and covers all protected identities such

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 11, November 2020 ||

as races, genders, nationalities, and generations. The non-disabled people are considered as temporarily able-bodied. Hate speech will be more common for disabled people than the ablebodied people. Hate speech on disables is due to the perception of disability by the violator but not due to actual disability of a person. There are several structural barriers for denying parental rights legally and removing sexual freedom as sexual autonomy on disabled people.
Intellectually disabled women are more vulnerable to violence at home. An able-bodied man will establish a relationship with the woman of an intellectual disability, initially, start being pleasant and gradually moves towards controlling her [61]. Even though the disabled persons are more vulnerable to hate violence, the hate reporting mechanism are less and not appropriate than other protected characteristics like gender/race. To maintain the social dignity of the disabled people, the local governments are required to have proper crime reporting and controlling systems.

3.5. Hybrid hate Speech

This category of hate speech is not related to a particular type. The hatred expressed in this form may be against more than one community and identity. That is the targets of a same anti-religion harassment may be Hindus and Muslims. A terrorist attack is one of the antecedent/parental trigger events for production and dissemination of hate on online social media like Twitter. Following an attack, the hate speech will be more at the time of impact stage, will start to reduce at inventory stage and will vanish during reaction stage (Williams and Burnap, 2015). Big data plays an important role in making policy and decision. A machine learning classifier is developed to recognize hate speech through twitter data following the Lee Rigby's murder incident. Generally, a combination of words as n-gram produces better results [21]. The learnability of the classifier depends on the set of features used to train. There is a necessity to improve overall performance by increasing classification accuracy, changing parameters and optimal kernel functions.

III. PROPOSED ALGORITHM

Proposed processing sequence is as follows:
Step I: Load the Dataset.
Step II: Pre-process the Dataset. For preprocessing following task is performed:

- Remove Noises.
- Tokenize the data.
- Tag the Tokens.
- Recognize the name entities.
- Lemmatization of Tokens.

Step III: Embed the tokens into Text Vectors and POS vectors.
Step IV: Apply Proposed Classifier.
Step V: End.

A) Data Collection: Data is collected from offensive language identification datasets.
Therefore to collect our data we have to use the same principle with hashtags on Twitter. More precisely, we will make some research to choose several initial hashtags which were likely to be related to hate speech. We will then make queries to collect tweets containing those initial hashtags. For each tweet collected, we looked at the other hashtags it contained. We then used those new hashtags to search for the next tweets. We will do this several times in order to find more specific hateful hashtags and thus increase the ratio of hate speech among all the tweets containing them. In this way, we could also find hashtags that couldn't have been found by other means because they are too specific.

B) Coupling hashtags with swear terms: In order to have a heuristic to evaluate the level of hate speech of a tweet and thus determine which **hashtag** should be kept for the data sampling, we have used a dictionary of hate words (also named hate base) found on the Internet. More precisely, we kept track of the number of hate words occurrences in the collected tweets. Each time we found a hate word in a tweet, we increased this number of occurrences for all the hashtags it contained.
We will also record the number of occurrences of every hashtag in the tweets. At the end, we only kept the hashtags with the highest ratios of hate speech to generate the next samples. We also decided not to keep the hashtags which themselves contained hate words. It should also be noted that the hashtags which had a high ratio of hate speech but which appeared only in a small number of tweets should be handled separately.
For example, a hashtag with ratio 1 could be a hashtag found in only one tweet containing a hate word. We decided to leave those cases aside because they weren't meaningful.

C) Cleaning the tweets and the hate words: To compare the words of the tweets with the words of the hate base we needed to tokenize the tweets. In order to tokenize a tweet and more specifically the hashtags, we separated the words of that tweet with spaces/punctuation (or group of punctuations). We also chose to separate each upper case letter (or

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 11, November 2020 ||

group of upper case letters) followed by lower case letters as different words. We then converted every word to lower case and did the same trick with the words of the hate base.

Let's consider for example the following tweet:

Stand! Fight! Win! Founders wrote #2A for self protection.

Europe should demand right to bear arms!! #Trump #London Attacks #MAGA, which was separated like this:

'stand', '!', 'fight', '!', 'win', '!', 'founders', 'wrote', '#', '2', 'a', 'for', 'self',

'protection', '.', 'europe', 'should', 'demand', 'right', 'to', 'bear', 'arms', '!!', '#',

'trump', '#', 'london', 'attacks', '#', 'maga'.

D) Hate word weighting: We noticed that the hate words from the hate base or more generally from any dictionary found on the Internet were not always associated with hate speech. Indeed some of them appeared more frequently than others in different contexts but not always with hate speech and therefore they advantaged the tweets containing them. To balance it, we decided to weight the hate words to give the less frequent hate words more importance. Consequently, when hate words which did not appear often were found, they were given a higher weight in order to compete with the hate words appearing more often.

3.3 Flow-Chart of Proposed Method

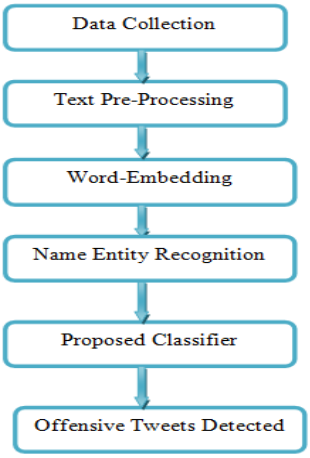


Figure 3.2: Steps for creating random forest.

IV. SIMULATION RESULTS

In this paper we are predicting that a sentence or word is hate word or not. We have a dataset that contains both hate word and not hate word. Here we are predicting that if a word is hate, then it will be classified as "Hate Class Word", and if t is not hate word, than it will be classified as "Not hate Word" class.

We have tried different advanced models with different pre-processing modules. For the first step of data cleaning, awaring the problem of previously excluding too much so called noise, we tried another way of data cleaning besides noise_cancelation called noise_replacement. While noise_cancelation directly deletes all the usernames mentioned in text, we reduce duplicate usernames into a single one. The intuition behind is that even though usernames appear in large part of the data, they are still unique in some circumstances and thus shall not be deleted directly. For the tokenizing part, we have tried the normal tokenizer and the tokenizer designed for tweets. Both of the tokenizers are built-in tools within NLTK. For the part-of-speech module, we have tried the spaCy tagger and the NLTK tagger. For the name entity tagger, we choose to use one built by spaCy. A detailed description of pre-processing pipelines for different models as well as their performance is shown below.

Evaluations of various classifier algorithms according to precision recall and f1-score are displayed below:

Table 5.1: Performance Evaluation:

Method	Precision (%)	Recall	F-score
Support Vector Machine	81.78	87.91	49.99
Logistic Regression	83.50	80.41	58.18
Ensemble Classifier	87.52	89.58	65.11
Proposed Method	**88.85**	**90.41**	**67.90**

It is observed that proposed classifier achieved best accuracy.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 11, November 2020 ||

V. CONCLUSION AND FUTURE WORK

The degree of hate associated with a hashtag was measured by the ratio between the number of hateful tweets and the total number of tweets - both containing the said hashtag. This approach proved to be appropriate. Indeed, among the hashtags our method highlighted, were "good" hashtags. A hashtag is considered "good" if the majority of the tweets mentioning it can be considered hateful by a human -after reading the said tweets.

Despite all the precautions taken in our proposed method research, we have nonetheless observed that the hate ratios of these "good" hashtags sometimes were inferior to other "less good" hashtags.

REFERENCES

[1] R. P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Comput. Survey. 51 (4) (2018) 85:1–85:30.

[2] N. Fandos, K. Roose, Facebook Identifies an Active Political Influence Campaign Using Fake Accounts, https://www.nytimes.com/2018/07/31/us/politics/ facebook-political-campaign-midterms.html, [Online; accessed 26-January-2019] (2018).

[3] L. Delisle, A. Kalaitzis, K. Majewski, A. de Berker, M. Marin, J. Cornebise, A large-scale crowd-sourced analysis of abuse against women journalists and politicians on twitter.

[4] P. Badjatiya, S. Gupta, M. Gupta, V. Verma, Deep learning for hate speech detection in tweets, in: Proceedings of the 26th International Conference on World Wide Web Companion, International World Wide Web Conferences Steering Committee, 2017, pp. 759–760.

[5] T. Grondahl, L. Pajola, M. Juuti, M. Conti, N. Asokan, All you need is “love”: Evading hate speech detection, in: Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security, CCS 2018, Toronto, ON, Canada, October 19, 2018, 2018, pp. 2–12.

[6] S. Agrawal, A. Awekar, Deep learning for detecting cyberbullying across multiple social media platforms, in: Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26- 29, 2018, Proceedings, 2018, pp. 141–153.

[7] Z. Waseem, Are you a racist or am i seeing things? Annotator influence on hate speech detection on twitter, in: Proceedings of the first workshop on NLP and computational social science, 2016, pp. 138–142.

[8] Almagor, R. C. (2011). Fighting Hate and Bigotry on the Internet. *Policy & Internet*, 3(3), 1-28, DOI: 10.2202/1944-2866.1059.

[9] E. Papegnies, V. Labatut, R. Dufour, G. Linares, Graph-based features for automatic online abuse detection, in: International Conference on Statistical Language and Speech Processing, Springer, 2017, pp. 70–81.

[10] S. Agrawal, A. Awekar, Deep learning for detecting cyberbullying across multiple social media platforms, in: Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings, 2018, pp. 141–153. Doi: 10.1007/978-3-319-76941-7_11.

[11] P. Saha, B. Mathew, P. Goyal, A. Mukherjee, Hate monitors: Language agnostic abuse detection in social media, CoRR abs/1909.12642. ArXiv: 1909.12642.

[12] M. Bojkovsky, M. Pikuliak, STUFIIT at semeval-2019 task 5: Multilingual hate speech detection on twitter with MUSE and elmo embeddings, in: J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, S. M. Mohammad (Eds.), Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6- 7, 2019, Association for Computational Linguistics, 2019, pp. 464–468. doi:10.18653/v1/s19-2082.

[13] S. Agrawal, A. Awekar, Deep learning for detecting cyberbullying across multiple social media platforms, in: Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings, 2018, pp. 141–153. Doi: 10.1007/978-3-319-76941-7_11.

[14] M. Dadvar, K. Eckert, Cyberbullying detection in social networks using deep learning based models; A reproducibility study, CoRR abs/1812.08046. ArXiv: 1812.08046.

[15] T. Grondahl, L. Pajola, M. Juuti, M. Conti, N. Asokan, All you need is “love”: Evading hate speech detection, in: Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security, CCS 2018, Toronto, ON, Canada, October 19, 2018, 2018,pp. 2–12.

[16] Seglow, J. (2016). Hate Speech, Dignity and Self-Respect. Ethical Theory Moral Practice, 19, 1103–1116, DOI: 10.1007/s10677-016-9744-3.

[17] Simons, R.N. “Addressing Gender-Based Harassment in Social Networks: A Call to Action”, In iConference 2015 Proceedings, http://hdl.handle.net/2142/73743.

[18] Edwards, S. S., “Cyber-Grooming Young Women for Terrorist Activity: Dominant and Subjugated Explanatory Narratives”, In Cybercrime, Organized Crime, and Societal Responses (pp. 23-46). Springer, Cham. ECHR-European Convention on Human Rights. (1950). http://www.echr.coe.int/Documents/Convention_ENG.pdf, Accessed 06 February 2017.

[19] Chua, V., Mathews, M., & Loh, Y. C. (2016). Social capital in Singapore: Gender differences, ethnic hierarchies, and their intersection. Social Networks, 47,138–150, DOI: 10.1016/j.socnet.2016.06.004.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 11, November 2020 ||

[20] Beckman, L., Hagquist, C., & Hellstrom, L. (2013). Discrepant gender patterns for cyberbullying and traditional bullying –An analysis of Swedish adolescent data. Computers in Human Behavior, 29, 1896–1903, DOI:10.1016/j.chb.2013.03.010.

[21] Burnap, P., & Williams, M. L. (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, DOI: 10.1002/poi3.85.

INNO SPACE

SJIF Scientific Journal Impact Factor

Impact Factor:
7.488

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details