Detection of Hate Speech in Social Media Using Machine Learning

Dr. Shubhangi D C, Professor, Visvesvaraya Technological University,
Department of Computer Science & Engineering VTU Regional Centre Kalaburagi-585105,Karnataka.
Ambika,Master of Technology, Visvesvaraya Technological University,
Department of Computer Science & Engineering VTU Regional Centre Kalaburagi-585105, Karnataka

Abstract

Humanity has benefited greatly from the increased usage of social media and knowledge sharing. However, this has resulted in a number of issues, including the spread and dissemination of hate speech messages. As a result, subsequent studies used a number of feature engineering techniques and machine learning algorithms to automatically detect this increasing issue in social networking platforms. To the best of our knowledge, there hasn't been a study that compares a number of feature engineering techniques and machine learning algorithms to see which one outperforms on a common publically available dataset. As a result, the purpose of this work is to compare the results of three feature engineering strategies and eight machine learning techniques. The bigram features outperformed the support vector machine algorithm by 79 percent total accuracy in the experiments. Our research has real-world implications and can be used as a benchmark in the field of recognizing hate speech automatically. Moreover, the output of different comparisons will be used as state-of-art techniques to compare future researches for existing automated text classification techniques.

Key Words: Hate Speech, Community Detection, NLP, Social Media

1. INTRODUCTION

1.1 Introduction

Hate speech has been more prevalent in recent years, both in person and online. Hateful content is being bred and propagated on social media and other internet platforms, which eventually leads to hate crimes.New Delhi: The Supreme Court on issued notices to the Centre and Twitter India on a plea seeking a mechanism to prevent spread of "fake news, hate news or anti-India posts" on the social media platform.

A three-judge bench led by Chief Justice SA Bobde issued the notices on a petition filed by BJP's Vinit Goenka saying hundreds of fake Twitter handles and bogus Facebook accounts in the name of eminent people were used to mislead people.

The notices come in the midst of a row between the government and Twitter over the latter’s refusal to clamp down on certain handles "with Khalistani and Pakistani links " which were used to post content on the protests against the new farm laws.

Twitter has cited freedom of speech rules to avoid clamping down on handles of journalists, politicians and activists though it has blocked many of the handles it was asked to block. The government of India has been unhappy with the limited compliance

of its order.

More than 10 per cent of the 35 million twitter handles and 350 million Facebook accounts in India were fake, the plea said, alleging they were used to peddle "hate speech and fake news" and cause social unrest and riots including the recent one in Delhi.

There should be a law under which action can be initiated against Twitter and its representatives in India for abetting such content, it said.

The plea filed through advocate Ashwini Dubey also sought that Twitter share its logic and algorithms with the Indian authorities for screening "anti-India tweets" and that all social media handles be made traceable,

2. Related Work

The most crucial step in the software development process is conducting a literature review. It is vital to determine the time factor, economy, and team strength before building the tool. After these requirements have been met, the next stage is to choose which operating system and programming language will be utilised to construct the tool.Once the programmers begin working on the tool, they will require a great deal of outside assistance. Senior programmers, books, and websites can all provide this assistance. Before building the system the above consideration are taken into account for developing the proposed system. We have to analyse Machine learning and Hate Speech Predicational analysis: With the rise of Social Media internet users became able to easily express and share theiropinionsaboutcompanies,products,services,event setc.Thus,companiesbecameinterested in monitoring what people say about their brands in order to get feedback orenhancetheirmarketingefforts.

In the field of social media monitoring, machine learning has a number of intriguing applications. It is employed to assess user opinions and categorise them as good, negative, or neutral (Also known as Hate Speech Prediction Analysis). It can also be used to determine whether the postings are objective or subjective, the natural language used in the posts, and whether the posts were written by the author.

[1] Dinakar et al. (2012), Sood et al. (2012b) and Gitari et al. (2015) follow a multistep approach, in which a classifier dedicated to detect negative polarity is applied prior to the classifier specifically checking for evidence of hate speech.

[2] Gitari et al. (2015) run an additional classifier that weeds out non-subjective sentences prior to the aforementioned polarity classification.

[3] Van Hee et al. (2015) use as features the number of positive, negative, and neutral words (according to a sentiment lexicon) occurring in a given comment text.

2.1 Proposed System:

To classify content as hate speech and predict whether a given text is hate speech or not, the proposed solutions used a variety of feature engineering techniques and machine learning algorithms. Algorithms handle training, testing on a dataset, and calculating accuracy.

3Methodology

Implement a high-accuracy system for automatically classifying tweets into Positive, Negative, or Neutral categories. The tweets can be compiled to form a summary of the overall Hate Speech Prediction on a given topic.

3.1.1 Tarining Data:

Training data should consist of tweets and its Hate Speech Prediction. In our case we have csv file withtwocolumns. First column isHate Speech Prediction and second column istweet.

3.1.2 Features:

All the symbols, URL are removed and word starting with number and word withoutHate Speech Prediction is removed. The remaining list of word is called features which is list of wordthat havesomeHate Speech Predictiont.

3.1.3 Machine Learning Algorithm:

Machine learning is the term for the algorithm that is used to train the classifier. It creates a classifier that has been trained. Multinomial Naive Bayes, Bernoulli Naive Bayes, Logistic Regression, Stochastic Gradient Descent, Support vector clustering, and Linear Support vector clustering are six other methods we have in our scenario.

Multinomial Naïve Bayes uses term frequency i.e. the number of times a given **term appears in** a document. ... After normalization, term frequency can be used to compute maximum likelihood estimates based on the training data to estimate the conditional probability.

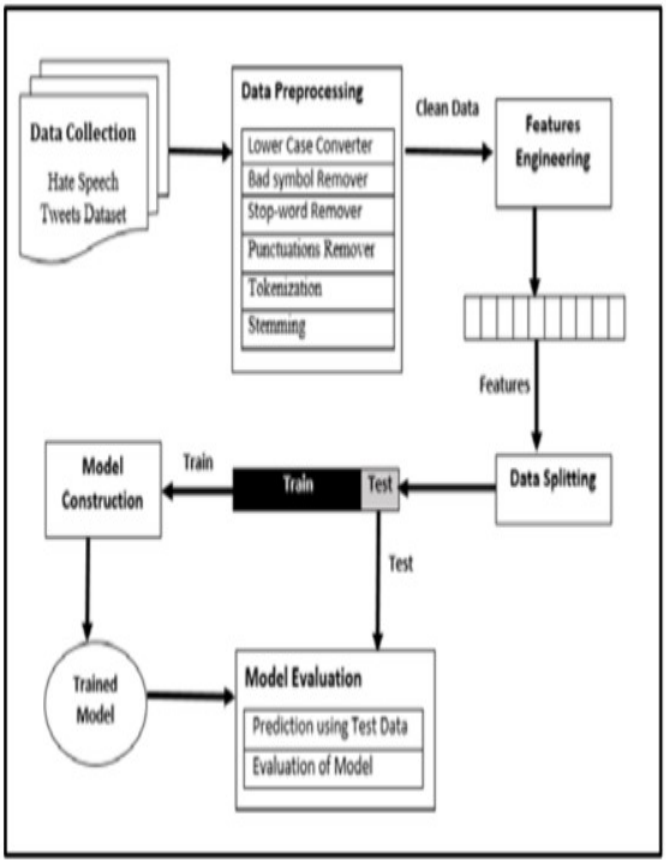
P(A|B) = P(B|A) P(A)
P(B)

Logistic Regression was used in the biological sciences in early twentieth century. It was then used in many social science applications. Logistic Regression is used when the dependent variable(target) is categorical.**Logistic Regression Equation is**

$$\ln\left(\frac{P}{1-P}\right)=b0+b1*x1+b2*x2\ldots\ldots+Bk*Xk$$

Stochastic gradient descentis an iterative method for optimizing an objective function with suitable smoothness properties. It can be regarded as a stochastic approximation of gradient descent optimization, since it replaces the actual gradient by an estimate thereof.

1. Initialize w:=0m−1,b:=0w:=0m−1,b:=0
2. for epoch e∈[1,...,E]e∈[1,...,E]:
   - 2.1. shuffle DD to prevent cycles
   - 2.2. for i∈[1,...,m]i∈[1,...,m] (where mm is the minibatch size):
     - 2.2.1. draw random example **without** replacement: ⟨x[i], y[i]⟩∈D⟨x[i],y[i]⟩∈D
   - 2.3. compute loss L:=1m∑mi=1L(y^[i],y[i])L:=1m∑i=1mL(y^[i],y[i])
   - 2.4. compute gradients Δw:=−∇Lw,Δb:=−∂L∂bΔw:=−∇Lw,Δb:=−∂L∂b
   - 2.5. update parameters w:=w+Δw,b:=+Δbw:=w+Δw, b:=+Δb



3.1.4 Input:

Theinputtoprogram issentenceortweet.

3.1.5 Classification Module:

We have seven classifier model. One for each machine learning algorithm that is used toclassifyinput andgives output as label.

3.1.6 Label:

LabelistheHate Speech Prediction outputgeneratedbyclassifier forthegiveninput.

4. Results

Hate Speech Prediction output from seven classifier can be used for voting to give which is the bestHate Speech Prediction.TheHate Speech Predictionwiththemajority votewins.Forexample,4outofsevenclassifier gives positive as output and other 3 as negative or neutral then confidence iscalculated as follows:

No. of positive = 4, No. of negative = 3

No.ofneutral=0, Output=Max (No.ofpositive,No.ofnegative,No.of neutral)

Output= 4 (positive), Confidence Level=(output\Total) * 100 %, ConfidenceLevel =(4\7) *100%=57.14%

Hencetheoutputwill bepositivewith 57.14confidence.

5. Conclusion

Hate Speech Prediction analysis has a wide range of applications. It's used to track customer reviews, survey replies, rivals, humanoid robots, and other things in social media monitoring and Voice of Customers (VOC). However, it can also be useful in business analytics and text analysis settings.

Hate Speech Prediction Analysis is

excellent, but it's a challenging assignment, as we all know, because every coin has two sides. The difficulty rises in tandem with the intricacy of the conveyed viewpoints. Some domains, such as product reviews, facial recognition, and span filters, are very simple, whereas others, such as novels, movies, art, and music, require more indirect expressions of opinion.

The beauty of social media Hate Speech Prediction analysis is that you're not searching for a needle in a haystack. Hate Speech Prediction mining examines big groups of people and trends. It means that with the raw amount of data, you can account for some fuzziness in Hate Speech Prediction classification; otherwise, we'll discover that the trends we're looking for aren't prevalent.

References

[1] Dinakar et al. (2012), Sood et al. (2012b) and Gitari et al. (2015) follow a multistep approach, in which a classifier dedicated to detect negative polarity is applied prior to the classifier specifically checking for evidence of hate speech.

[2] Gitari et al. (2015) run an additional classifier that weeds out non-subjective sentences prior to the aforementioned polarity classification.

[3] Van Hee et al. (2015) use as features the number of positive, negative, and neutral words (according to a sentiment lexicon) occurring in a given comment text.

[4] P. Mell and T. Grance, "The NIST definition of cloud computing," *Nat.Inst. Stand. Technol.*, vol. 53, no. 6, pp. 50–50, 2009.

[5] H. T. Dinh, C. Lee, D. Niyato, and P. Wang, "A survey of mobile cloud computing: Architecture, applications, and approaches," *Wireless Commun. Mobile Comput.*, vol. 13, no. 18, pp. 1587–1611, 2013.

[6] J. Chase, R. Kaewpuang, W. Yonggang, and D. Niyato, "Joint virtual machine and bandwidth allocation in software defined network (sdn) and cloud computing environments," in *Proc. IEEE Int. Conf. Commun.*, 2014, pp. 2969–2974.

[7] H. Li, W. Sun, F. Li, and B. Wang, "Secure and privacy-preserving data storage service in public cloud," *J. Comput. Res. Develop.*, vol. 51, no. 7, pp. 1397–1409, 2014.

[8] Y. Li, T.Wang, G.Wang, J. Liang, and H. Chen, "Efficient data collection in sensor-cloud system with multiple mobile sinks," in *Proc. Adv. Serv. Comput., 10th Asia-Pac. Serv. Comput. Conf.*, 2016, pp. 130–143.

[9] L. Xiao, Q. Li, and J. Liu, "Survey on secure cloud storage," *J. Data Acquis. Process.*, vol. 31, no. 3, pp. 464–472, 2016.

[10] Hate Speech PredictionAnalysisonTwitter throughTopic-basedLexicon ExpansionbyZhixinZhou, XiuzhenZhang,andMark Sanderson

[11] TwitterHate Speech PredictionAnalysis:TheGoodtheBadandtheOMG!By EfthymiosKouloumpis, TheresaWilson, JohannaMoore

[12] ModelingandRepresentingNegationinData-drivenMachineLearning-basedHate Speech Prediction AnalysisbyRobert Remus.

[13] Hate Speech PredictionAnalysisofTwitterDataUsingMachineLearningApproachesandSemantic Analysis byGeetikaGautam.

.