A Logistic Regression Model for Hate Speech Recognition

Sania Zehra[1*], Faraz Doja[2*]
{ sania.znaqvi@gmail.com[1], farazdoja@jamiahamdard.ac.in[2]}

Department of Computer Science, Jamia Hamdard, Block D, Hamdard Nagar, New Delhi

Abstract: People feel entitled to express their opinions on social media, and many of these opinions may be hostile to others. Hate speech recognition is essential in today's world, where it is important to keep track of what is posted in public for all to see. This paper discusses ways to detect hateful language on any topic using data mining, natural language processing, and machine learning techniques on the social media site Twitter. A Logistic Regression Model for Hate Speech Recognition is proposed. The research conducted provides details of all the hate commenters as well as an overview of the topics on which the hate is being projected the most.

Keywords: Data Mining, Natural Language Processing, Machine Learning, Supervised Learning, Logistic Regression

I Introduction

With the exponential growth and exposure to the virtual world of social media, modern techniques are required to keep a check on what is being posted online. Various social media platforms are continuously working on this major issue- social media hate, which influences the world in a way never envisioned. Regulation and detection of this hate speech language is possible using techniques like natural language processing and machine learning. The data is plenty but is unprocessed, this huge heap of data can be processed in ways which can make social media a safer place for all. Various machine learning algorithms are used to process and analyse the data and make predictions based on either labelled or unlabelled data.

II Related Work

Hajime W., et al. (2018) proposed a method which detects hate speech patterns and most common unigrams automatically and uses these, along with sentimental and semantic features, to classify tweets as hateful, offensive, or clean. Gitari et al. (2015) proposed a method for extracting sentences from some of the most popular hate sites in the United States. They classified the sentences into three categories: strongly hateful (SH), weakly hateful (WH), and non-hateful (NH). They used semantic and grammatical pattern features, ran the classification on a test set, and got an F1-score of 65.12%. Kwok and Wang (2013) proposed a method for

ICIDSSD 2022, March 24-25, New Delhi, India

DOI 10.4108/eai.24-3-2022.2318769

detecting racist tweets directed at black people. They used unigram features, which resulted in a 76% accuracy for the binary classification task. Salminen, et al. (2018) proposed a method that combined hate/non-hate classification with additional classes (e.g., accusation, humiliation, etc.). They proposed SVM with a linear kernel. Nobata, et al. (2016) proposed a method for performing the classification task into two classes using lexicon features, n-gram features, linguistic features, syntactic features, pretrained features, word2vec features, and comment2vec features, and obtained an accuracy of 90%. Burnap and Ohtsuki (2015) distinguished hate speech utterances from clean speech utterances using typed dependencies (i.e., the relationship between words) and bag of words (BoW) features. T. Mandl et al. (2019) proposed a method for detecting hate speech and offensive content in Indo-European languages. Alexander Brown (2017) proposed a method for determining what makes online hate speech different from offline hate speech. Zhang Z. et al. (2019) suggested Deep Neural Network architectures that act as feature extractors and are particularly successful at capturing hate speech meanings. On Twitter's largest collection of hate speech datasets, methods are tested, and they are demonstrated to perform up to 5% better than the top approach in macro-average F1 or 8% better in the more complex situation of identifying hateful content. Ziqi Z., et al. (2018) proposed a paper that introduces a new method based on a deep neural network that combines convolutional and gated recurrent networks. On the largest collection of publicly available Twitter datasets to date, they conduct a thorough evaluation of the method against a number of baselines and the state of the art. They demonstrate that, when compared to previously reported results on these datasets, our proposed method is able to capture both word sequence and order information in short texts, and it sets new benchmarks by outperforming on 6 out of 7 datasets by 1 to 13% in F1. They also introduce a new dataset that covers a wide range of issues to the collection of datasets already used for this endeavour. Joseph M. (2009) gives an overview of all logistic models including binary, proportional, ordered, partially ordered, and unordered categorical response regression procedures. Other topics discussed include panel, survey, skewed, penalized and exact logistic models.

III Machine Learning

Computers may learn without being explicitly programmed thanks to a field of computer science called machine learning (ML). As the name suggests, the machine processes the data in a human-like way, and unlike traditional programming, the results may or may not be provided to the machine for analysis. It is a name for a collection of algorithms that make intelligent predictions based on a set of data. These data sets are huge, with millions of distinct data points. ML has recently achieved what looks to be a human level of semantic understanding and information extraction, as well as the capacity to discern abstract patterns more accurately than human specialists. ML is a type of artificial intelligence that has the potential to change the world in the twenty-first century. Recent advances in its underlying architecture and algorithms, as well as the development in the size of datasets, have resulted in increased computer competence in a variety of sectors. These include things like driving a car, translating languages, using chatbots, and performing better than humans at sophisticated board games like

Go. Due to massive volumes of data, exponential development in computer power and breakthroughs in algorithm design modern ML has evolved as a formidable tool driven by the needs of the web industries. Today, a wide range of ML methods, referred to as models are in use[1]. The qualities of the data as well as the type of intended outcome influence the model selection for a given problem. There are 4 types of ML Algorithms:

1. *Supervised Learning* – An ML algorithm, with a known dataset that comprises intended inputs and outputs, and must figure out how to get at those inputs and outputs.
2. *Unsupervised Learning* – with no known dataset, the ML algorithm analyses data to find patterns.
3. *Semi-supervised Learning* – use both labelled and unlabelled data
4. *Reinforcement Learning* – focuses on structured learning procedures in which an ML algorithm is given a set of actions, end values and parameters[2].

Numerous algorithms focus on function approximation issues when the task is embodied in a function (for example, output a "hate" or "not hate" label given an input transaction). The goal of the learning process is to increase the function's accuracy through experience gained from a sample of known input-output pairings. When a function is formed by an initial search, factorization, optimization, or simulation, it can sometimes be explicitly represented as a parameterized functional form. Training is the process of determining out the optimum settings for these parameters to enhance the performance metric. Even if a function is implicit, it is typically dependent on parameters or other degrees of freedom that may be controlled. For the training of the ML model for this research a supervised learning algorithm called logistic regression is used. Logistic regression is a method for modelling the probability of a discrete result given an input variable. Most common logistic regression models have a binary result, such as true or false, yes or no and so on. Modelling scenarios with more than two discrete outcomes with multinomial logistic regression is possible. In classification jobs, logistic regression is a valuable analysis method for assessing if a new sample fits best into a category. Logistic regression is a classification model and not a regression model despite its name. Logistic regression is a simple and effective method for binary and linear classification. It's a simple classification model that produces outstanding results with linearly separable classes. It is a widely used categorising method in the industrial world[4]. A logistic function is used to forecast the categorical dependent variable of interest after data is entered into a linear regression model. In supervised learning, the operator gives the ML algorithm a dataset which is known, with desired inputs and outputs, and the system figures out how to get those inputs and outputs, which is just what we want to do to train our model. While we know the correct answers to the problem, the algorithm analyses patterns in data and learns from observations to provide predictions. The operator rectifies the algorithm's predictions and the process is repeated until the algorithm achieves a high degree of accuracy. The following concepts are a part of supervised learning:

1. *Classification*: In classification tasks, the ML algorithm must draw a conclusion from observed values and decide which group fresh observations belong to. When categorising tweets as 'hate' or 'non-hate,' for example, the software must consider existing observational data and filter the tweets accordingly.
2. *Regression:* It focuses on a single dependant variable and a set of other changing variables, making it particularly effective for forecasting and prediction.
3. *Forecasting*: It is the process of creating future predictions based on historical and current data, and it is frequently used to analyse trends[5].

By estimating probabilities using logistic regression model, it is possible to understand the relationship between the dependent variable and one or more independent variables. This form of analysis can assist you in predicting the chances of an occurrence or a decision occurring. The training data set contains binary values, i.e. '0' and '1', where '0' depicts a non-hate tweet and '1' detects hate tweet and the coefficients of the logistic function are determined using the TF-IDF vectors using equation(1).

$$f(x) = \frac{1}{[1+exp(-(b+a_1\,x_1+\cdots a_n x_n)]} \quad (1)$$

IV Proposed Methodology

Data analysis is the process of analysing, cleaning, modifying, and modelling data with the goal of identifying useful information, generating conclusions, and supporting decision-making. Data mining, also referred to as knowledge discovery in data (KDD), is the process of extracting patterns and other important information from large data sets[6]. The adoption of data mining techniques has skyrocketed in recent decades as a result of the development of data warehousing technology and the emergence of big data, assisting organisations in transforming unstructured data into insightful knowledge. Professionals continue to struggle with scalability and automation difficulties despite the fact that technology is constantly improving to handle enormous amounts of data. Data mining has improved corporate decision-making through clever data analytics. These studies' data mining methods fall into one of two categories: either they describe the target dataset or forecast results using machine learning (ML) algorithms. These techniques are used to organise and filter data, providing the most useful information, from fraud detection to user patterns, inefficiencies, and even security breaches. Data mining is a data analysis approach that is defined as a method for extracting useable data from a bigger quantity of raw data[7]. For this research, the data mining has been done through 'Tweepy', a python library for accessing the Twitter API. A Twitter developer account provided all the necessary access codes for accessing the Twitter data through 'Tweepy'. Once the API call was set up, the data was mined for the trending hashtags in India and then tweets on general topics using hashtag recognition as well as general hashtags were saved in a comma-separated values(CSV) file. A data set is also collected from Kaggle in order to analyse the hate speech. This data is labelled; that is, the hate tweets are marked as "1" and the non-hate tweets are

marked as "0" (Fig. 1.). This data set will be used for the training process. For testing, the data can be collected either on a specific topic or general tweets. This testing data is collected (Fig. 2.) and then stored in a CSV file. There are approximately 22500 non-hate speech tweets and around 2500 hate speech tweets in the training dataset.

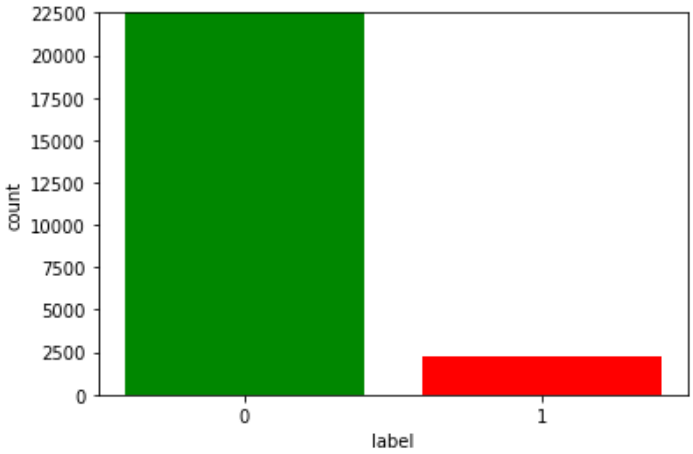

Fig. 1. Graph of the data used for analysis (0-Non Hate Speech, 1- Hate Speech)

	id	tweet
0	31963	#studiolife #aislife #requires #passion #dedic...
1	31964	@user #white #supremacists want everyone to s...
2	31965	safe ways to heal your #acne!! #altwaystohe...
3	31966	is the hp and the cursed child book up for res...
4	31967	3rd #bihday to my amazing, hilarious #nephew...
...	...	...
17192	49155	thought factory: left-right polarisation! #tru...
17193	49156	feeling like a mermaid ð #hairflip #neverre...
17194	49157	#hillary #campaigned today in #ohio((omg)) &am...
17195	49158	happy, at work conference: right mindset leads...
17196	49159	my song "so glad" free download! #shoegaze ...

Fig. 2. Raw testing data

Once the two data sets are ready, they are concatenated into a single data frame for cleaning and natural language processing (NLP). It is then cleaned for analysis using the Natural Language Toolkit (NLTK). After removing special characters, numbers, hash tags and words with less

than 2 characters, the most frequently used words are detected and printed, along with the total word count. All the stop words are also removed from the data. Once the tweets are cleaned and the most frequently used words are detected, they are converted into numbers. For this process, the TF-IDF vectorization is used. The TF-IDF score is an element (numeric) that portrays the importance of a phrase to each word in a document. The TF-IDF score is calculated using equation(2).

*TF (term frequency) * IDF (inverse-document frequency)*

$$\frac{n(frequency\ of\ term)}{N(number\ of\ all\ words)} * log_{10} \frac{D(number\ of\ sentences\ in\ document)}{d(number\ of\ sentences\ containing\ that\ term} \quad (2)$$

Logistic regression model is used in a program that could classify hate speech. Scikit-learn is a library in Python that has a highly optimized version of logistic regression implementation, supports multiclass classification task. It has been used for implementing logistic regression on the training dataset. The labelled hate speech tweets will determine the coefficients of the logistic function using the TF-IDF vectors. The textual data from the tweets is transformed into a word cloud, which is then utilized to display keyword metadata. The textual data collected from the tweets is converted to a word cloud, used to depict keyword metadata to visualize free form text. The flowchart (Fig. 3.) represents all the steps for hate speech recognition.

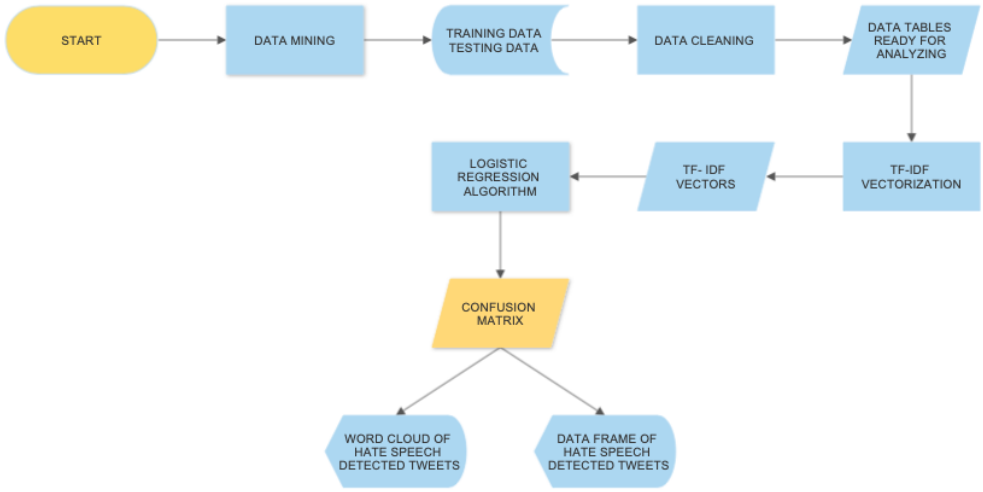

Fig. 3. Steps for Hate Speech Recognition

The hate tweets are displayed using word cloud (Fig. 4.) and the user ids and hate tweets were put into a data frame (Fig. 5.). The importance of each word is depicted by the font size and

colour in the word cloud. After analysis, the model scored an accuracy of 93%, which is determined using the function model.score().

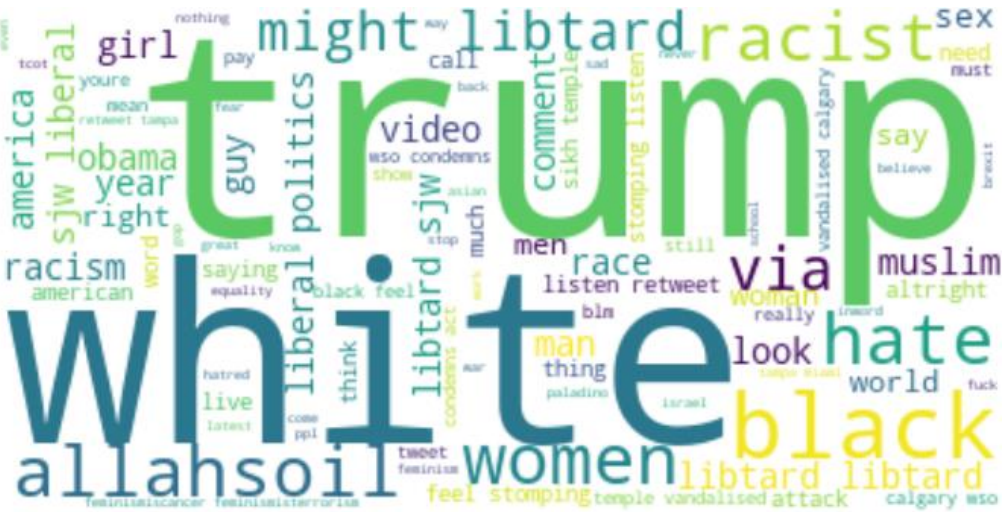

Fig. 4. Word Cloud of the hate speech detected tweets

	IDs	Hate Tweets
0	22377	samsunggalaxys rooster simulation climb vast e...
1	22398	facemask run risk looking unintentionally more...
2	22406	teen difference men women sex
3	22426	situation fallujah iraq dire city siege via
4	22440	make show called buddysfamilyvacation gorge fo...
...	...	...
668	31890	feels know gods grace mercy infinite tho chape...
669	31905	model
670	31914	chateaubriand stovells absolutely lovely food yum
671	31937	know live things make less things make unhappy...
672	31945	youre surrounded even deserve yet hateful

Fig. 5. Data Frame of the Hate Speech Detected tweets with the User IDs

V Conclusion

The research conducted shows how ML and data analysis techniques can be used for various purposes; it extracts useful information of given raw data and the model used for this research can be used in detecting cyber hate. As ML algorithms gain experience, their accuracy and efficiency improve. This enables them to make more informed decisions. These algorithms learn to generate more accurate predictions faster as the amount of data grows. The sample twitter data consists of thousands of tweets and running analysis on it shows people often vent on social media about things that might enrage them. The logistic regression ML model was successful at detecting hate tweets and gave an overview of the topics at which most of the hate comments were projected. Judging from the analysis, most of the hate comments are targeted towards politics, gender, race, and religion. These comments if detected with accuracy can help reduce social media driven violence and possibly civil wars where people encouraging and influencing virulent content can be detected by the governments and actions could be taken against them. The analysis would be helpful in detecting the user IDs of people with a trend or pattern of posting hate comments and they could be banned from Twitter to make it a safe place for all. However, this analysis if done with malevolent intent, might be used by dictators or fascists for oppression of minorities and anyone who might oppose them, putting the safety and privacy of many at risk.

VI Limitations

As the data used for analyses is labelled by the researcher, it is biased and it depends on an individual whether to label a comment or word as hateful or not. The data used for training the model has an unbalanced ratio of hate- to-non-hate tweets. As a result, despite achieving a 93% accuracy rate, the model is unable to recognise all hate speech in the testing data. This analysis was performed on pre-mined testing and training data, while the model is capable of running on real-time data as well. For real-time mining, more resources and necessary infrastructure is required, which would give better results. If there are less data than features, the logistic regression should not be implemented because this could lead to overfitting. Also, the assumption of linearity between the dependent and independent variables is a major limitation of logistic regression. It not only indicates the suitability of a predictor (coefficient size), but also the direction of relationship that can be positive or negative. When there are several or non-linear decision boundaries, logistic regression tends to underperform.

References

[1] James A. Nichols, Hsien W. Herbert Chan and Matthew A. B. Baker- Machine learning: applications of artificial intelligence to imaging and diagnosis (11 Feb, 2019)
[2] M. I. Jordan and T. M. Mitchell - Machine learning: Trends, perspectives, and prospects (17 July, 2015)
[3] Machine Learning Series: Regression-4 (Logistic Regression) – Arun (Aug 16, 2021)
[4] Towards data science: Logistic Regression — Detailed Overview (March 15, 2018)
[5] SAS- A guide to the types of machine learning algorithms and their applications-By Katrina Wakefield, Marketing, SAS UK
[6] Thomas W. Edgar, David O. Manz - Logistic Regression (2017)
[7] IBM Cloud Education: What is Data Mining? (15 January, 2021)