Me Too Movement Sentiment Analysis

DM. Rama Bai, Charishma Kuna, J. Sreedevi, G. Shantha

Abstract**:** ***Sentiment Analysis (SA) is a current field of study in text mining. The subjectivity of text, sentiment, and opinions are treated computationally by SA. This study examines the sentiment of the tweets containing "#metoo". As a comparison, the same analysis was performed on the MenToo movement. MeToo started picking up significance in India with the expanding ubiquity of the global development, and later gathered sharp force in October 2018 in the film business of Bollywood, focused in Mumbai, when Tanushree Dutta blamed Nana Patekar for lewd behavior. An Indian filmmaker has joined calls for the development of a "#MenToo" movement for men's rights, saying it should be "as important as #MeToo. This case study gathers around 20,000 tweets from the major cities of India for the duration of a week. Tweets were analyzed through the 'sentiments' dataset of tidytext (afinn, bing, nrc) and RSentiments dataset. The goal was to understand the overall sentiment better and find the associated patterns. With the hashtag analysis, it can be seen that #metoo was associated with the film industry, whereas #mentoo was more rooted in the cause. The comparison of likes and retweets shows that the #metoo movement has over 70% more engagement than #mentoo.***

Keywords**:** ***#metoo, #mentoo, Sentiment Analysis, RSentiments, TidyText.***

I. INTRODUCTION

Sentiment analysis has been an important research area of data mining for the last 20 years. Their peer reviews influence the interests of human beings. So, whenever a decision has to be made, people often pursue other's reviews to get a general opinion. Critique sites, forum discussions, blogs, microblogs, and social media digital platforms provide a platform for reviews. Although every website has a vast quantity of reviews, the ordinary human reader will have a problem in ascertaining relevant sites, mining and summarizing the reviews which often results in a situation that the right decision could not be reached.

This problem applies to a person as well as for associations, organizations, ideological groups, and so on.

That is the reason mechanized savvy opinion examination frameworks, which can precisely give the general supposition and other related data in less time (than if done physically), is essential in the present information-driven world.

Men and women share experiences of sexual assault to provide comradeship to survivors and to shed light on how underreported these cases are, this movement emerged to be the #metoo movement on Twitter. This movement engulfed India when Bollywood actress Tanushree Dutta came forward with an accusation against Nana Patekar for sexual harassment. This incident shifted the paradigm of the movement, giving way for millions of women to share their voices. The internet was filled with mentions of #MeToo stories, shedding light on a taboo subject in India.

Revised Manuscript Received on April 30, 2020.

***** Correspondence Author

Dr M Rama Bai*, Professor, Department of Computer Science and Engineering, Mahatma Gandhi Institute of Technology, Hyderabad, India. Email: vallapu.rama@gmail.com, mrmabai_cse@mgit.ac.in.

Ms.Charishma Kuna, Student, Department of Computer Science and Engineering, Mahatma Gandhi Institute of Technology, Hyderabad, India. Email: charishmak98@gmail.com

Mrs. J. Sreedevi, Assistant Professor**,** Department of Computer Science and Engineering, Mahatma Gandhi Institute of Technology, Hyderabad, India Email: jasthysreedevi@gmail.com

Dr G. Shantha, Assistant Professor, Department of Mathematics & Humanities, Mahatma Gandhi Institute of Technology, Hyderabad, India. Email: gshantha_maths@mgit.ac.in

II. LITERATURE SURVEY

Rachana Bandana (2018) et al. [1] "Sentiment Analysis of Movie Reviews Using Heterogeneous Features." "Rachana Bandana has proposed a methodology, heterogeneous features, for example, AI-based and Lexicon based features and administered ML learning methodologies like Naive Bayes (NB) and Linear Support Vector Machine (LSVM), to fabricate the framework model. From execution and perception, it can be inferred that utilizing proposed heterogeneous features and hybrid approach can get an exact assessment investigation framework contrasted with other pattern frameworks. In the future, for comprehensive information, these heterogeneous features can be utilized for building progressed and progressively precise models utilizing Deep Learning (DL) algorithms."

Chae Won Park (2018) et al. [3] "Sentiment Analysis of Twitter Corpus Related to Artificial Intelligence Assistants" "Better experience is one of a huge ebb and flow issues in the user's research. A procedure that improves the client's experience ought to be required to assess the ease of use and emotion. Most importantly, assumption examination dependent on client's conclusions can be utilized to comprehend client's propensity. This paper means to make a rule what artificial intelligence right hand is factually better."

Metin Bilgin (2017) et al. [4] "Sentiment Analysis on Twitter data with Semi-Supervised Doc2Vec". "Emotions are dissected on the messages shared on Twitter with the goal that clients' thoughts on the items and organizations can be resolved. Assumption investigation causes organizations to improve their items and administrations dependent on the criticism acquired from the clients through Twitter. In this examination, it was meant to perform sentiment analysis on Turkish and English Twitter messages utilizing Doc2Vec.

The Doc2Vec algorithm was run on Positive, Negative, and Neutral labeled information utilizing the Semi-Supervised learning strategy, and the outcomes were recorded."

Alyssa Evans (2018) et al. [5] "#MeToo: A Study on Sexual Assault as Reported in the New York Times." "This research looks at the degree to which inclusion by The New York Times of the #MeToo development incorporates a differing foundation of casualties of sexual assault and provocation. Source portrayal in media impacts the public's impression of social issues and gatherings. This contextual investigation tracks statistic inclusion of lewd behavior and ambush in a prominent news association. Information assembled looks at The New York Times surrounding of unfortunate casualties and inclusivity of announcing over a two-month term in 2017."

Ana Tarano (2017) et al. [6] "Tracking #metoo on Twitter to Predict Engagement in the Movement." "The objective of this task was to all the more likely to comprehend and anticipate what sorts of tweets get especially high consideration and commitment. In doing as such, knowledge into the potential arrive at future internet-based life developments by understanding what content is probably going to contact the vast majority. Inspected 3,750 tweets inside the #metoo development; by looking at the word events inside the contents of the tweets, the option to anticipate whether a tweet would be retweeted over a mean edge with 90% exactness is analyzed."

III. PROPOSED APPROACH

The dataset is created from extracting tweets by creating a Twitter developers account, then getting a standard search API. Using the API, the respective code can request the tweets satisfying the constraints specified in the code. It is stored in the file format Comma-separated values (CSV)n in which there are columns with the text of the tweet, a number of retweets, whether it has been favorited or not, among other useful information. The dataset contains around 20,000 tweets equally divided among MeToo and MenToo topics. To process a large amount of data for mining, structuring of the text data is an important step. Well-structured data will enable us to obtain useful information by applying machine learning algorithms. This is a critical step; if not done correctly, the output obtained will be erratic. The objective is to expel characters less pertinent to discover the sentiment of tweets, for example, punctuation, unique characters, numbers, and terms that don't convey much weightage in setting to the content.

Features are used primarily for eliminating noise, improve classification accuracy, and to reduce vocabulary size. Heterogeneous features are produced using a machine learning algorithm like a bag of words, TF-IDF, etc.

Term frequency-inverse document frequency (TF-IDF): The associated weight is often used in text mining plus data retrieval. This statistical measure (weight) is used to assess the weightage of a word in a document. The significance is straightforwardly corresponding to the number of times a word comes up in the record; however, for the recurrence of the word in the corpus, it is not the situation. Varieties of this scheme are often utilized by web crawlers as a critical tool in recording and placing a report's pertinence dependent on a client inquiry.

Bag of words: The BOW model is a plain description utilized in natural language processing and information retrieval (IR). Methodology for Sentiment Analysis (SA) is given in below Figure 1

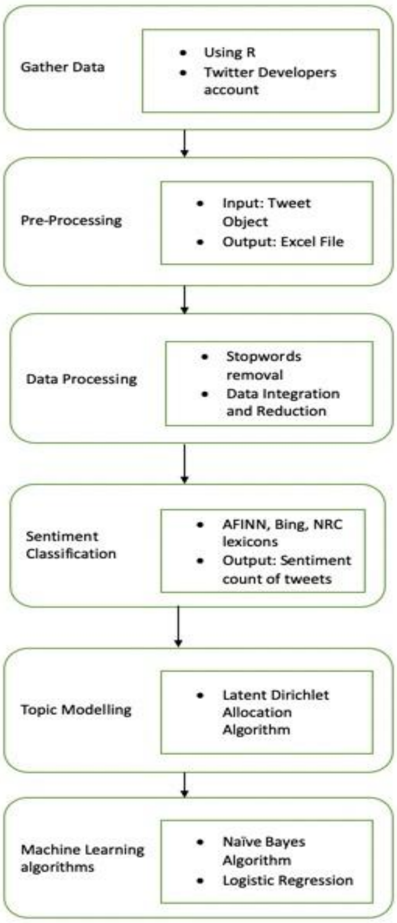

Figure 1: Methodology for Sentiment Analysis

Sentiment Analysis uses sentiment polarity for categorizing tweets in text. Tweet text having two contradictory reviews is called polarity. This polarization is positive or negative or neutral. Machine learning algorithms like Logistic Regression (LR) and Naive Bayes (NB) are used to organize and learn text into positive and negative categories in the word level.

Logistic Regression is a machine learning classification algorithm, which is used to estimate the probability of a categorical dependent variable. A binary variable, either 1 (yes) or 0 (no) can be the output or the target variable in logistic regression. The concept of logistic regression predicts, P(Y=1) as a function of X, in other words.

Logistic Regression is one amongst the foremost in style ways in which to suit models for categorical information, particularly for binary response information in information Modeling. It's the most necessary (and maybe most ordinarily used) member of a model category called generalized linear models. Unlike linear regression, logistic regression will directly predict probabilities (values that are restricted to the (0,1) interval); however, those probabilities are well-calibrated relative to the probabilities expected by another classifier, such as Naive Bayes.

Logistic regression preserves the training data’s marginal probabilities. The model’s coefficients also provide some clue about the relative significance of every input variable.

If the dependent variable (target) is categorical, logistic regression is used. Consider the case of tumor. The target variable could be malignant (1) or non-malignant (0).

Consider a scenario within which the goal is to work out whether an email is spam. If we are using linear regression for this problem, a threshold must be set, depending on that classification can be made. If the actual class is malignant, the predicted continuous value 0.4 and the threshold value is 0.5, then the data point will be marked as non-malignant, which can result in serious real-time consequences. From this instance it may be concluded that linear regression is not appropriate for the problem of classification. Linear regression is unbounded and this gives an illustration of logistic regression. Its meaning is purely from 0 to 1.

Logistic regression is generally used where Binary or Dichotomous is the dependent variable. That means that the dependent variable can only take two possible values like "Yes or No," "Default or No Default," "Living or Dead," "Respondent or Non-Respondent," etc. Independent factors or variables can be categorical or numerical variables.

Naive Bayes works on the principle of conditional probability given in the below equation and it is based on Bayes theorem. It is a supervised machine learning algorithm commonly used for binary classification.

$$P(A|B) = (P(B|A) * P(A)) / P(B) \qquad (1)$$

The above equation is the representation of Bayes theorem with two events A and B. The occurrence of event A is given by P(A) and same applies for P(B) and P(B|A) is the conditional probability that event B occurs, given that A has occurred.

Consider ' k ' attributes with potential true or false values, therefore we need a scale 2^k joint distribution table. To train the classifier we will have to train all these values which is not feasible when ‘k’ is large.

Naive Bayes uses the assumption of conditional dependence to reduce the problem. It is this assumption that makes the Bayes Theorem naive. Instead of attempting to compute the values of each attribute value, consider here P (d1 d2 d3|h), they are assumed to be conditionally independent given the target value and calculated as P(d1|h) * P(d2|H) and so on. Consider that your data has the phrase “Mango Tea”, Naive Bayes treats both words as separate attributes that is “mango” alone and “tea” alone, even if there exists something called “mango tea”. This is a very strong assumption given the type of data that exists but Naive Bayes still manages to perform well in the task of classification.

One of the mainstream approaches to look into the opinion of a document is to consider the content as a blend of its individual words and, in this manner, the sentiment of the full content as the expansion of the opinion score of the individual words. This is the most utilized methodology in Sentiment Analysis, which takes advantage of well-established eco-system. A specific range of strategies and dictionaries exist for evaluating the sentiment or feeling in the document. In sentiments dataset, the tidytext package has many sentiment lexicons, the three frequently used lexicons are afinn, bing, and nrc.

For sentiment analysis, various classes are categorized in a binary manner yes/no by the nrc lexicon. These classes could be positive, negative, anticipation, disgust, anger, and many such classifications. The bing lexicon classifies words in a binary way into positive and negative groups. The AFINN lexicon allows words with a rating that ranges between -5 and 5, with positive values specifying positive opinion and negative values specifying negative opinion. Model approach applied to the text data is given in below Figure 2

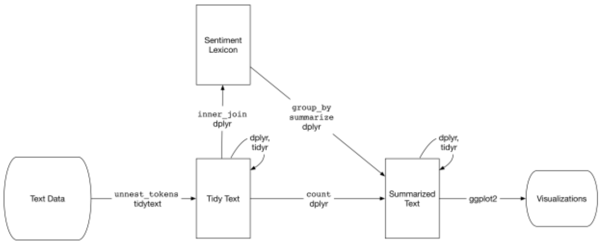

Figure 2: Model approach applied to text data

Topic modeling is a technique for the unsupervised arrangement of text documents that discovers a characteristic method of things, such as clustering on numeric information, in any event, when we are uncertain about what we are searching for.

Latent Dirichlet allocation (LDA) is an undoubtedly well-liked method for fitting a topic model. It regards each report as a blend of subjects, and each subject as a mix of words. This enables reports to "overlap" each other as far as content is concerned, rather than being isolated into independent groups, in a technique that mirrors conventional utilization of regular language.

Naive Bayes is a collection of computations bolstered by applying Bayes hypothesis with a ground-breaking (naive) supposition, that each component is autonomous of the others, to conjecture the classification of an instance. They are probabilistic classifiers. This manner will figure the likelihood of every classification utilizing the Bayes hypothesis, and the class with the most probability likelihood will be produced. Naive Bayes classifiers are, with favorable results, applied to a few areas, especially Natural Language Processing (NLP).

In the early twentieth century, Logistic Regression was employed in the biological sciences. It was employed in several social science applications. It is utilized when the dependent variable (target) is categorical.

IV. CONCLUSION

By plotting the top 5 hashtags used in the tweets of MeToo and MenToo, it can be concluded that most of the MeToo hashtags were related to the film industry compared to MenToo, which was more rooted in the cause. The resulted statistic is given below in Figure 3

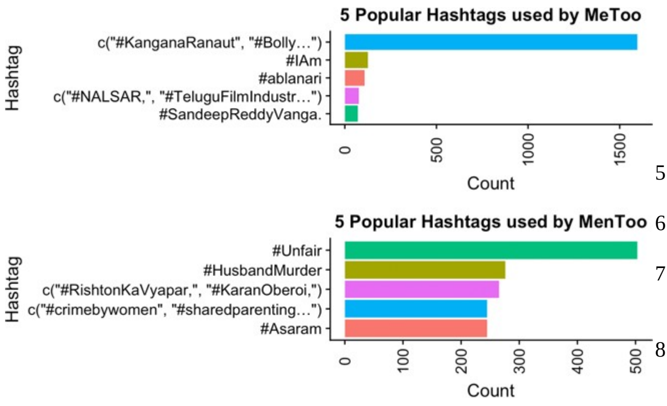

Figure 4.1 Hashtag Analysis
Figure 3: Hashtag Analysis

The Likes Ratio and Retweet Ratio statistic are shown below in Figure 4 and Figure 5 respectively. To gauge the engagement received by #MeToo movement, the Likes count and the Retweet count was measured and compared against #MenToo movement. It is clearly seen that #MeToo has more engagement than #MenToo.

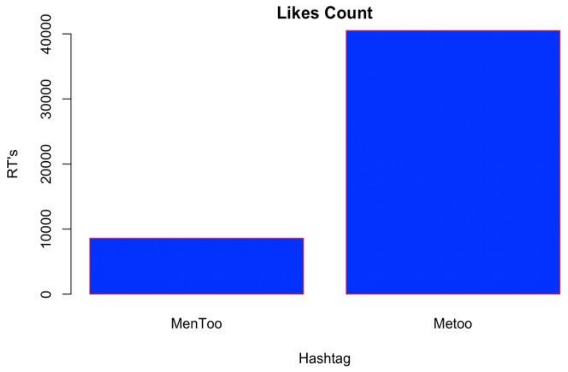

Figure 4: Likes Ratio

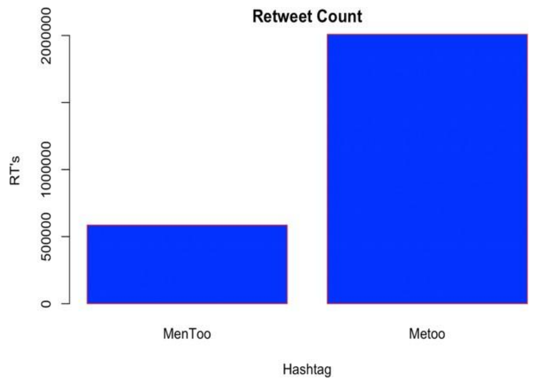

Figure 5: Retweet Ratio

If there were an access to a larger dataset, the region-wise analysis would give a more in-depth understanding. It would also help formulate the POSH (Prevention of Sexual Harassment) policy.

REFERENCES

1. Rachana Bandana, “Sentiment Analysis of Movie Reviews using Heterogeneous Features,” 2018 2nd International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech)
2. R. Bais and P. Odek, “Sentiment Classification on Steam Reviews,”Stanford University, 2017
3. Chae Won Park, “Sentiment analysis of Twitter corpus related to artificial intelligence assistants,” 2018 5th International Conference on Industrial Engineering and Applications (ICIEA).
4. Metin Bilgin, “Sentiment analysis on Twitter data with semi-supervised Doc2Vec,” 2017 International Conference on Computer Science and Engineering (UBMK).
5. Alyssa Evans, “#MeToo: A Study on Sexual Assault as Reported in the New York Times,” Occam’s Razor, Vol. 8, Article 3.
6. Ana Tarano, “Tracking #metoo on Twitter to Predict Engagement in the Movement,” Stanford University, 2017
7. Mäntylä, M.V., Graziotin, D., Kuutila M.: The Evolution of Sentiment Analysis - A Review of Research Topics, Venues, and Top Cited Papers. (arXiv:1612.01556v1 [cs.CL]) (2016)
8. Stray J.: What do Journalists do with Documents? Field Notes for Language Processing Research. Stanford Journalism Department, https://journalism.stanford.edu/cj2016/files/What%20do%20journalists%20do%20with%20documents.pdf (2016)
9. Ninth International Conference on Language Resources and Evaluation, Reykjavik, May 26-31 (2014).′Abbasi, A., Hassan, A., Dhar, M.: Benchmarking Twitter Sentiment Analysis Tools. In: Calzolari N., Choukri K. (eds) LREC
10. Liu B.: Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, San Rafael, CA (2012)
11. Tumasjan A., Sprenger T.O., Sander P.G. et al.: Predicting Elections with Twitter:What 10: 4th International conference on Weblogs and Social Media, Washington, D.C., May 23-26 (2010)′140 Characters Reveal about Political Sentiment. In: Hearst M. (ed) AAAI
12. Liu B.: Sentiment Analysis and Subjectivity. In: Indurkhya N, Damerau FJ (eds) Handbook of Natural Language Processing, 2nd edn. Chapman & Hall, Boca Raton, FL, p. 627-661 (2010)