# Me Too Movement Sentiment Analysis

Charishma Kuna[1], Dr. M. Rama Bai[2]
[1]*charishmak98@gmail.com*, [2]*rama@mgit.ac.in*

### *Abstract*

*Sentiment Analysis (SA) is a current field of study in text mining. 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 creation 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 better understand the overall sentiment and find the associated patterns. With the hashtag analysis, it can be seen that #metoo was associated with the film industry where as #mentoo was more rooted to the cause. The comparison of likes and retweets shows that #metoo movement has over 70% more engagement than #mentoo*

***Index Terms:*** *Sentiment Analysis, social media, #metoo, #mentoo, India, TidyText, RSentiments*

## 1 Introduction:

Sentiment analysis has been an important research area of data mining for the last 20 years. The interests of human beings are influenced by their peers' reviews. So, whenever a decision has to be made, people often seek out other's review to get a general opinion. Critiques sites, forum discussions, blogs, microblogs, and social media digital platforms provide a platform for reviews. Normal human pursuer will experience difficulty in recognizing pertinent locales, removing and abstracting the audits so the correct choice can't be come to. Although every website contains a vast amount of reviews, the average human reader will have trouble in identifying relevant sites, extracting and abstracting the reviews so the right decision cannot be reached. This 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 truly necessary 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 a 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.

## 2 Literature survey

Rachana Bandana (2018) et.al. [1] **“Sentiment Analysis of Movie Reviews Using Heterogeneous Features”**. “In the proposed 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) used to fabricate the framework model. From execution and perception, infer that utilizing proposed heterogeneous features and hybrid approach can get an exact assessment investigation framework contrasted with other pattern frameworks. In future for large information, we can utilize these heterogeneous features 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 the most 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 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.”

## 3 Proposed Approach

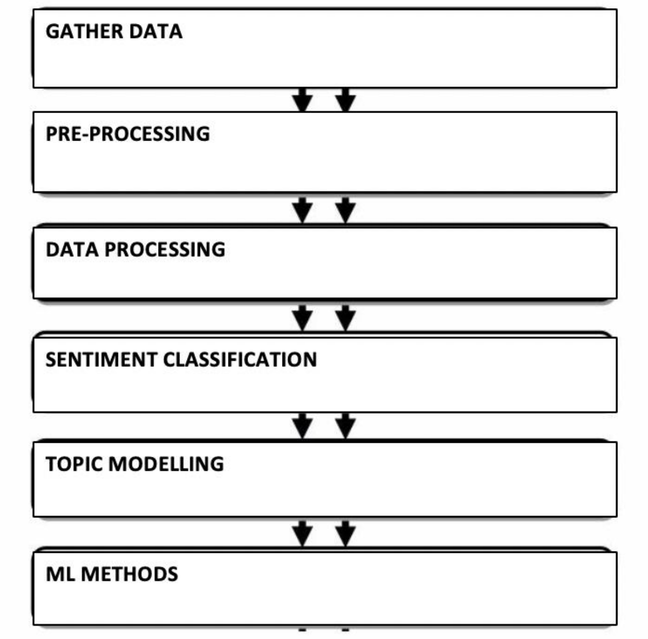


**Fig 3.1** Methodology for SA

### 3.1 Twitter Dataset

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, 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.

### 3.2 Preprocessing

To process large amount of data for mining, structuring of the text data is an important step. Well-structured data will enable to obtain useful information by applying machine learning algorithms. This is a critical step, if not done properly, 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 which don't convey much weightage in setting to the content.

### 3.3 Features

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

Term frequency-inverse document frequency (TF-IDF): The associated weight is often used in text mining and information 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 shows up in the record however it is not the situation for the recurrence of the word in the corpus. Varieties of this scheme are frequently utilized via web crawlers as a critical tool in scoring and positioning 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).

### 3.4 Classification Algorithm

Sentiment Analysis uses sentiment polarity for categorizing tweets in text. Tweet text having two contradictory reviews is called polarity. This polarity can be 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.

### 3.5 Sentiment analysis

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 well established eco-system. Certain range of strategies and dictionaries exist for evaluating the sentiment or feeling in document. The tidytext package contains several sentiment lexicons in the sentiments dataset.

The three general-purpose lexicons are

- AFINN
- bing, and
- nrc

Classes of positive, negative, anticipation, disgust, anger, fear, surprise, joy, sadness and trust are categorized from words in a binary manner (“yes”,”no”) by the NRC lexicon. The bing lexicon classifies words in a binary manner into positive and negative groups. The AFINN lexicon allots words with a rating that ranges between -5 and 5, with positive values specifying positive opinion and negative values specifying negative opinion.

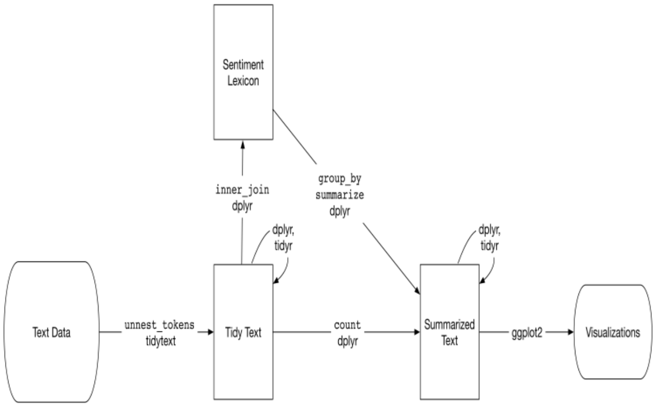


**Figure 3.2** Model approach applied on text data

### 3.6 Topic modelling

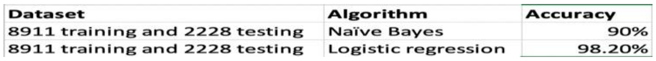
Topic modeling is a technique for unsupervised arrangement of text documents that discovers characteristic arrangement of things, such as clustering on numeric information, in any event, when we are uncertain about what we are searching for.

Latent Dirichlet allocation (LDA) is a certainly 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.

### Machine Learning models

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

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



| Dataset | Algorithm | Accuracy |
|---|---|---|
| 8911 training and 2228 testing | Naïve Bayes | 90% |
| 8911 training and 2228 testing | Logistic regression | 98.20% |

## 4 Conclusion

From experiments, it can be concluded that most of the MeToo hashtags were related to the film industry compared to MenToo which was more rooted to the cause. After performing the analysis of likes and retweets, MeToo has a substantially larger engagement than MenToo. If there was an access to a larger dataset, region wise analysis would give a more in depth understanding and 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