A Detailed Survey on Sentimental Analysis on Social Media

Gulbakshee Dharmale, Shreenidhi Karjagi, Harshal Kotalwar, and Shakyadeep Khobragade

Abstract Public opinion or point of view matters a lot as it plays vital role in shaping country's image all around the world, may it be on certain products, decisions or their own feelings/expressions. At this time, people are more expressive on social media, more rather than in person, they express or share their daily tasks their feelings, opinions some good or bad incidents, and many more on social media, and we can collect this data and process into classifiable form. The data can be collected through API and then transformed into classifiable form. To classify data, we will be foreseeing three algorithms which are Naïve Bayes, maximum entropy, and lexicon-based. The main reason of gathering this data is to process and generate positive and negative notions of certain demography or specific person. This data can be useful in many ways such as to check if a person is in good mental health condition or suffering through depression through his/her positive negative posts on social media. Multi-national companies can collect insights on their products based on peoples review and opinion on social media. MNC and e-commerce stores can also product recommendation system based on those reviews. In short, sentiment analysis plays a very vital role in our economy as well as health care.

Keywords Opinions · Positive · Negative · Social media · Sentiments · Naïve Bayes · Maximum entropy · Lexicon-based · API's · Algorithms · Classifiers

1 Introduction

Social media comprehends of billions of users across the world; on this platform each and every topic such as current affairs economy, sports, reviews, opinions, and much more is discussed. In this discussion, the outcome is either positive or negative,

G. Dharmale · S. Karjagi (✉) · H. Kotalwar · S. Khobragade
Department of Information Technology, Pimpri Chinchwad College of Engineering, Pune, Maharashtra, India
e-mail: shreenidhikarjagi29@gmail.com

G. Dharmale
e-mail: gulbakshi.dharmale@pccoepune.org

V. K. Gunjan and J. M. Zurada (eds.), *Proceedings of 3rd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications*, Lecture Notes in Networks and Systems 540, https://doi.org/10.1007/978-981-19-6088-8_22

for example, if Indian cricket team wins, Indians will share positive or happy posts, but in case of rival team, it could be otherwise. Trillions of data are generated each day, and some these are personal expressions such as if a person is feeling sad, he/she shares that he/she is feeling sad today and vice-versa. People are tending to be more expressive on social media the reason for such expressiveness towards social media was found out to be that netizens (abbreviation for people using Internet), and they would not judge them based on their ideology unlike in our surroundings if the ideology of a person is not plausible, then the person and his opinions are mocked or make them feel embarrassed. We can use this data and conclude positive or negative notions from them for many purposes as discussed prior. Data can be gathered through API which are free for developers so that they can work on such projects and make the best use out of it, for classification, we have three possible algorithms Naïve Bayes, maximum entropy, and lexicon-based.

2 Literature Survey

David Osimo and Francesco Mureddu research challenge on opinion mining and sentiment analysis using [5] Naïve Bayes classifier had an accuracy of 78.80%.

Adobe Social Analytics: This assesses the impact of social media on businesses by determining how online conversations and communities impacts on marketing performance. It compares the influence of those interactions with important business dimensions such as revenue and brand value after capturing and interpreting the ongoing dialogues. Aside from that, it tracks how businesses interact with their clients on social media, such as how Facebook posts influence site traffic and purchase decisions [6].

A Literature Survey on Sentiment Analysis Techniques Involving Social Media and Online Platforms: This survey depicts the possible techniques and importance of sentiment analysis. This survey consists of all information required for sentiment analysis such as need of sentiment analysis the importance theoretical knowledge of sentiment analysis [7].

Adaptive Co-training SVM for Sentiment Classification on Tweets: X. Cheng and H. Shen compare both SVM and maximum entropy and state that SVM [8] has 77% accuracy, and maximum entropy has 85%.

Table 1 represents the prior research on sentiment analysis and their outcomes to analyse which approach is better.

3 Approaches

There are many possible approaches for analysing and classifying these approaches for analysing natural language processing which extracts opinions off the statement and sends it for classification. For classifier, we will be looking at three possible

Table 1 Literature survey

S. No	Title of paper	Author and year	Methodology	Result and accuracy
1	Sentiment analysis for social media	Jayasanka et al. [1]	In this research, the data was classified using Naïve Bayes classifier	The results show that the data collected had an accuracy of 78.80%
2	Improved lexicon-based sentiment analysis for social media analytic	Jurek et al. [5]	The approach in this research was lexicon-based data for classification of the data	The results depict that the accuracy of the positive negative values is 78%
3	Sentiment analysis on Twitter using maximum entropy and SVM	Ermatita et al. [4]	The classifier used in this research was SVM and maximum entropy, and best amongst both is chosen	The result has two outcomes that are SVM of accuracy 77%, and maximum entropy has accuracy of 85%
4	Naive Bayes and sentiment classification	Stanford university (2021)	Method used was Naïve Bayes for spam detection	The result given out was precision of 86% and accuracy of 85.43%

approaches, there are more than three approaches, but the accuracy level precision is not up to the mark, that is, the reason we will be foreseeing only three approaches those are as follows:

I. Naïve Bayes
II. Maximum entropy
III. Lexicon-based

3.1 Naïve Bayes

It is a supervised learning method based on the Bayes theorem that is used to solve classification issues. It is commonly used in text classification with a large training dataset. The Naive Bayes classifier is one of the simplest and most effective classification algorithms available, and it assists in the development of fast machine learning models capable of giving accurate predictions. It is a probabilistic classifier, which means it makes predictions based on an object's probability. It is an algorithm used to find most probable value for classifying test data in most appropriate category. In this case, test data is document tweets. There are 2 document stage.

I. Training
II. Classification

$$\text{General formula}: P(H|X) = P(X|H) \times P(H)/P(X) \tag{1}$$

where

a. $(H|X)$ is the probable final probability (rearward probability) of hypothesis H occurs when given evidence E occurs.
b. $P(X|H)$ is the probability that when E occurs, it will affect hypothesis H.
c. $P(H)$ is the prior probability hypothesis H occurs regardless of any evidence.
d. $P(X)$ is prior probability evidence E regardless of hypothesis or other evidence.

In this theory, 2 variables are used as characteristics as hypothesis (H) and sentiment as evidence (E). The other 3 variables will be used as metadata as of sentiment. Sentence containing of many words which is very challenging in practice to decide which one might be called aspect or feature it is presumed each word is an aspect or feature. Application of Bayes theorem (Fig. 1):

$$P(F|K) \times P(K)/P(F)$$

$F =$ Feature word
$K =$ Category/sentiment value because the features or words that support the same category can be various, there are features F_1, F_2, F_3 can be transformed into

$$P(K|F_1, F_2, F_3) = P(F_1, F_2, F_3) \times P(K)/P(F_1, F_2, F_3) \tag{2}$$

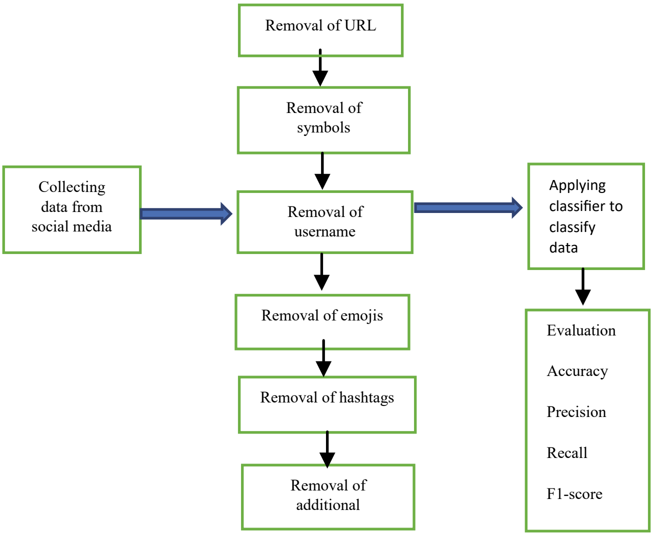


Fig. 1 Flowchart of classifier

It requires that the evidence (in this case is a feature or word) that exists is independent of each other, then formula can be changed to $P(K|F_1, F_2, F_3) = P(F_1|K) \times P(F_2|K) \times P(F_3|K) \times P(K)/P(F_1) \times P(F_2) \times P(F_3)$. If described in general, the formula is given as

$$P(K|F) = \prod P\,q\,i = (F_1/K)/P(F) \tag{3}$$

3.2 Maximum Entropy

It is a machine learning algorithm based on empirical data (information that comes from the research) and provides probabilities on which sentence belongs to specific class.

Unlike the Naive Bayes classifier, it does not assume that the data is independent; however, instead of using probabilities to define the model's parameters, it employs a search strategy to identify the set of parameters that will maximise the classifier's performance. The outputs of the procedure, like any learning technique, are dependent on the input dataset. There are no assumptions made about the features' relationships. The fundamental goal of the strategy is to optimise entropy in the system in order to predict label condition distribution in each class.

$$P_{\mathrm{me}}(c/b, \lambda) \frac{\exp\left[\sum \lambda_i f_i(c, b)\right]}{\sum_c \exp\left[\sum \lambda_i f_i(c, b)\right]} \tag{4}$$

where c is a class, b is social media like Twitter, λ is a weight of vector.

3.3 Lexicon-Based

This approach calculates sentiment function of whole document/data or set of sentences from semantic orientations of lexicons. The orientations can be positive negative or neutral. The dictionary of lexicons can be automatically or manually be generated, SentiWordNet dictionary is most commonly used. Firstly, lexicons are found from the dataset, and then, SentiWordNet can be used to discover the synonyms and antonyms to expand the dictionary. This technique uses adjectives and adverbs to discover the semantic inclination of text. For calculating inclination, the adjective and adverb union are extracted with their sentiment value. These can be transformed to single score for whole text.

So these are the three best approaches for classification in sentiment analysis; after classification, the process is not over we need to see how to import data how to make it classifiable (i.e. suitable for classification) then evaluate the data.

4 Procedure

The process of analysing basically consists of three phases.

1. Crawler/Web crawler
2. Sentiment analysis tool
3. Data mining (optional)
4. Evaluation

1. **Crawler**: The main motive of crawler is to collect data from Twitter or any social media platform to data source for easiness of classification and analysis. For importing, it has to use Twitter API to gain access to Twitter data. Twitter has many endpoints that have been customised for certain use cases like user stream, public stream, and site streams; for crawler, we will be using public stream.
2. **Sentiment analysis tool**: This is where the operations on data occur such as processing data into classifiable form, classifying data. The data here is same as the data that is crawled from the social media. It analyses sentiments and depicts it in accurate variance. It uses algorithms and techniques to derive most accurate output for given sentiments with the help of proper classifier. Further after retrieving, labelled text corpus is to extract features from it and to train the classifier. The entire system revolves around on how favourable this method of extraction is. Therefore, it uses several methods in sentiment analysis as follows.

   - Unigrams—Considers word individually in sentences as feature set of corresponding categories. In this case, it does not contemplate any relation between words.
   - Unigrams except stop words—It is similar to unigrams, besides it does not contemplate stop words that is a list of words which often appears in almost all sentences that has no context
   - Bigrams—Consider adjoining pair of words from sentences as a feature set of corresponding categories.
   - Bigrams except stop words—It is similar to bigrams feature set, besides from the words in the stop words list
   - Most informative unigrams and bigrams—Gets the feature set with unigrams and bigrams with most information and greatest frequency (occurrence of data).

Out of these methods, the most informative unigrams and bigrams are selected for data mining for forecasting future trends based on current analysed data.

3. **Data Mining**: During the product profiling, decision tree is used after correlating it with clustering technique, and for the trend analysis and forecasts it using Holt-Winters method, it is capable of analysing seasonal data and predict proper values for the future. (Holt-Winters method: it is a model of time series behaviour. Forecasting always requires a model.
4. **Evaluation**: After classification of data into positive negative or neutral sentiments, the most important task is evaluation. In evaluation process, we check for

accuracy precision and recall. The data which is more precise and accurate is preferable and suitable because the algorithm with more accuracy and precision is considered ideal. Operations performed in evaluation are

- Accuracy
- Precision
- Recall

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{5}$$

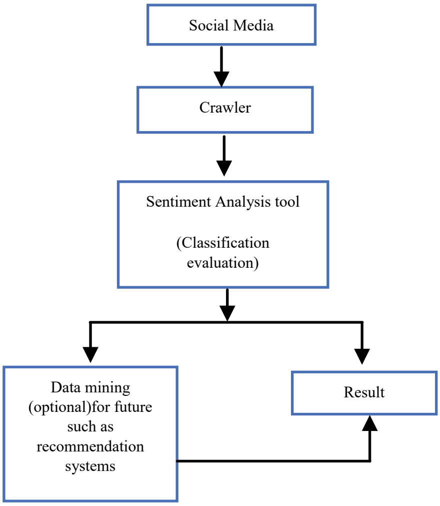
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

- Accuracy: It is used to determine which model is best at identifying relationship and patterns between variables.
- Precision: It defines number of positive class that actually belong to positive class.
- Recall: It states how many true states were found.

The basic structure of the model would be as shown in Fig. 2.

Fig. 2 Flowchart of model


Social Media
Crawler
Sentiment Analysis tool
(Classification evaluation)
Data mining (optional)for future such as recommendation systems
Result

5 Conclusion

Through this we can conclude that we have foreseen every successful feasible and best suited approaches for sentimental analysis. The main question is which is the best approach in case of Naïve Bayes approach the data or sentiments is assumed independent of each other which is quite difficult every word is related to each other, but still, the resultant has good accuracy and precision as it collects positive and negative words. In case of maximum entropy unlike Naïve Bayes, data is not assumed independent and consider a whole which would be good for accuracy, but it requires more time to train and also overfitting may occur. Lexicon-based approach is complete different than above two it searches for positive negative words extracts its sentiment scores through the help of SentiWordNet and evaluates the sentiment of the statement it calculates orientation of whole document which is time consuming, and the accuracy is low compared to two other approaches. Many researches have stated that both maximum entropy and Naïve Bayes give out same accuracy and precision, but maximum entropy has an issue of overfitting that may affect accuracy in further cases or datasets, so with this we can conclude that Naïve Bayes is the most preferable and suitable approach for sentimental analysis, and also the most important part is accuracy which depends on dataset, each data may have different accuracy; in some cases, maximum entropy may give more accuracy than Naïve Bayes, but in this case, Naïve Bayes performs better than Maximum Entropy.

References

1. Jayasanka SC, Madhushani MDT, Marcus ER, Aberathne, IAAU, Premaratne SC (2013) Sentiment analysis for social media. In: Conference: information technology research symposium, vol 4, Project: Market potential of thumba karawila, November 2013
2. Adobe social analytics US social analytics best practice guide (2016)
3. A literature survey on sentiment analysis techniques involving social media and online platforms. Int J Sci Technol Res 9(6):1–8
4. Ermatita E, Cindo M, dian Palupi Rini (2020) Sentiment analysis on Twitter using maximum entropy and SVM. Published 17 April 2020 Computer Science
5. Jurek A, Mulvenna MD, Bi Y (2015) Improved lexicon-based sentiment analysis for social media analytic. Sec Inform 4
6. Jurafsky D, Martin JH (2021) Naive Bayes and sentiment classification. In: Speech and language processing. Stanford University. Copyright © 2021. All rights reserved. Draft of December 29, 2021
7. Osimo D, Mureddu F (2012) Research challenge on opinion mining and sentiment analysis. Technical report, September 2012
8. Adaptive co-training SVM for sentiment classification on tweets. CIKM'13: Proceedings of the 22nd ACM international conference on information & knowledge management, October 2013, pp 2079–2088