ISSN Print: 2394-7500
ISSN Online: 2394-5869
Impact Factor: 8.4
IJAR 2021; 7(10): 391-394
www.allresearchjournal.com
Received: 16-08-2021
Accepted: 19-09-2021

Shruti Patil
Department of Computer Science, All India Shri Shivaji Memorial Society's College of Engineering, Savitribai Phule Pune University, Pune, Maharashtra, India

Siddhant Patil
Department of Computer Science, All India Shri Shivaji Memorial Society's College of Engineering, Savitribai Phule Pune University, Pune, Maharashtra, India

Corresponding Author:
Shruti Patil
Department of Computer Science, All India Shri Shivaji Memorial Society's College of Engineering, Savitribai Phule Pune University, Pune, Maharashtra, India

An overview of sentiment analysis of hashtags from social media

Shruti Patil and Siddhant Patil

DOI: https://doi.org/10.22271/allresearch.2021.v7.i10f.9078

Abstract
The use of social has been rapidly increasing since a decade. It is a well-known platform to create-share-promote your ideas and or business. The people from all over the world spend a lot of time on various social media platforms such as Twitter, Instagram and Facebook. There is massive number of photos, tweets, videos, etc. are posted every day. Therefore, this leads to numerous data coming right away from the users to the social media handlers. At some point, this data has to be analysed for betterment of the users and for many commercial uses. Nowadays, the most abundant amount of data comes from the hashtags that people use for many reasons such as growing their account, growing their business and so on. In short, people tend to express their feelings through social media with the help of hashtags. These hashtags are to be analysed in order to structure the data properly. An analysis of social media hashtags is one of the best deals for marketing strategy. It tells you how your products react with your customers and if needed, what are the possible changes that can be done in order to boost your business using hashtags. In this paper, we have done the study of analysing sentiments of words/phrases preceded by hashtag sign "#" taken from social media which includes determining the sentiment polarity of hashtags and its literal category as positive, negative or neutral.

Keywords: Sentiment analysis, sentiment analytics, lexicon-based analysis, natural language processing, hashtag analysis, data mining

Introduction
Sentiment analysis is a natural language processing technique which is used to determine whether the data or statement data or statement or word is positive, negative or neutral. Its main aim is to extract subjective information from texts or words in natural language in the form of sentiments or opinions. It is often useful for the business analysis as it helps to determine what the customer actually feels about their work or product. It is used to detect the sentiment on social media such as Twitter, Instagram, Facebook, etc. Due to the rapid development in digitalized economy, people from all over the world have started performing all basic tasks over the internet that they used to do normally or physically. The increase in online stores i.e., e-commerce websites for the products have made the people incline towards digitization. The online business has increased over the time. However, to increase the productiveness and quality of products, the handlers often ask for revies and rating. And then, they analyse these reviews so as to boost the business. Since there are millions of people who use these websites, there are millions of reviews coming every day. Hence, it becomes difficult for the people to analyse each and every review, and here the sentiment analysis comes into picture. With the help of sentiment analysis, one can take measures to maintain and improve their business in the market. Therefore, sentiment analysis plays a vital role in business.

There are several techniques used to classify the sentiment. The major three techniques are Machine learning approach, Lexicon based approach and Hybrid approach. Machine learning approach uses machine learning algorithm. Lexicon-based approach makes use of dictionary which is further explained in this paper. Hybrid approach makes use of both machine-learning and lexicon-based approach combined together. In this paper, we have done an empirical study on Lexicon based approach which is simplest method to extract sentiments of hashtag words from the tweets or posts from the public account of social media.

Lexicon-based approach

Lexicon-based approach is one the two main techniques used for the sentiment analysis of a sentence or words or phrases. It deals with lexicons or the internal dictionary of all the words present in the lexicon of a particular language. This dictionary can be created manually as well as automatically generated. In general, a piece of text or a word is represented as bag of words, and then the sentiment value such as positive, negative or neutral is assigned to it from the dictionary or bag of words. Then the sum or average function is used to calculate the total sentiment.

Lexicon-based approach deals with lexeme such as tokens or words. It splits the sentence into tokens and processes them. These words are classified as positive or negative or neutral opinions. In this paper, we are just focusing on the words or phrase preceding with a Hashtag sign "#". The words or word with "#" generally do not have any space between them, added we don't need to aggregate the total sentiment value.

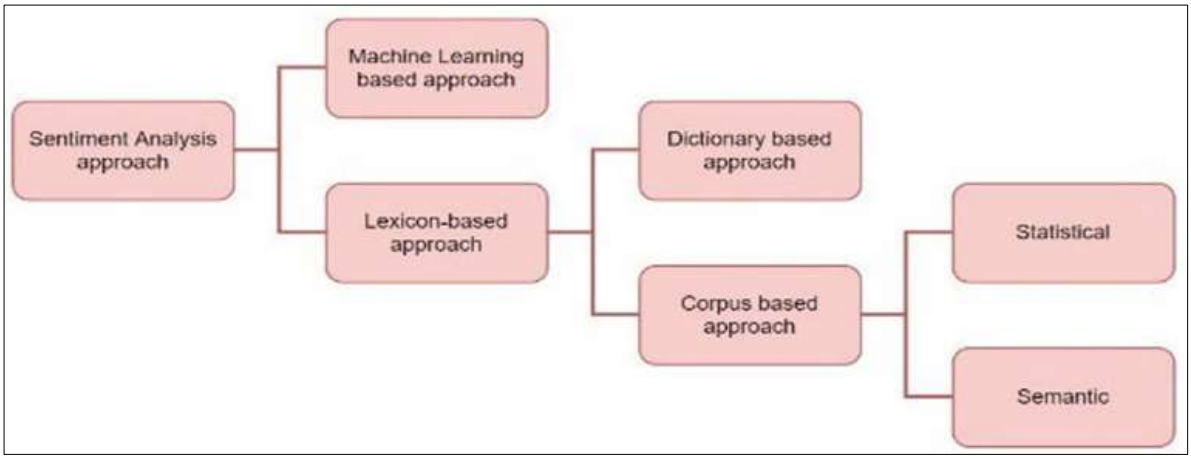

Fig 1: Sentiment analysis approaches

Lexicon based approach is classified as follows

1. Dictionary based approach

In this approach, a dictionary is created which consists of all the synonyms and antonyms of each word we use for analysis. This dictionary can be made manually; or it can be auto- generated.

2. Corpus based approach

In this approach, sentiments of context-specific words are analysed. It is less efficient than dictionary-based approach because large corpus of English language is needed to be made [7]. The methods of corpus-based approach are further classified into two categories:

Statistical approach: The word which show positive behaviour, have positive polarity. If it shows negative behaviour, it has negative polarity. If it shows both positive and negative frequency, it has neutral behaviour.

Semantic approach: It assigns sentiment values on the basis of antonyms and synonyms of the words from the dictionary. Basically, it is done by finding synonyms and antonyms with respect to the words that we are analysing.

Terminologies used in Twitter

- **Emoticons:** Emotion icon is the pictorial representation of facial expression using some special characters to express the feeling.
- **Target/URLs:** Twitter and other social media uses "@" to direct to targeted user. It also makes use of URLs like "https://..." to target certain page.
- **Hashtags:** Hashtags are basically used to mark or highlight certain topic. It is identified when the word is preceded by "#" symbol such as "#happy", "#singing", etc. And if a phrase uses hashtag, it is represented without any space such as #havingfun, "#sohappy", "#eatingpizza", etc.

Sentiment Lexicon

A sentiment lexicon is the set of words or the collection of words having the sentiment polarity. Each word in the lexicon is assigned a value either positive or negative in the range of -100 to 100 (-100 being most negative and 100 being most positive). For example, since we are dealing with hashtags,

1) *"I #love that movie!"* – The sentiment of this sentence is positive because of the word "love" which is according to the lexicon holds the positive polarity.
2) *"I #hate that movie*!" – The sentiment of this sentence is negative because of the word "hate" which according to the lexicon holds negative polarity.

Based on observation, in certain cases the polarity is neither positive nor negative and thus occurs with neutral polarity. For example, *"The movie was #Ok!"* – The sentiment of this sentence is neutral because of word "Ok" which holds neither positive nor negative polarity.

The probability (P) [8] formula to represent polarity is given by,

$$P(positive \mid w) \text{ --- } positive\ polarity$$
$$P(negative \mid w) \text{ --- } negative\ polarity \quad (1)$$

Above equations determines the probability that the random text containing this word is positive or negative based on the labelled data. This is important to decide the sentiment when the sentence contains mixed (positive and negative) sentiment.

General Process of Lexicon-based Sentiment analysis

There are 5 basic steps to follow in sentiment analysis of a word or a statement [6]. They are as follows,

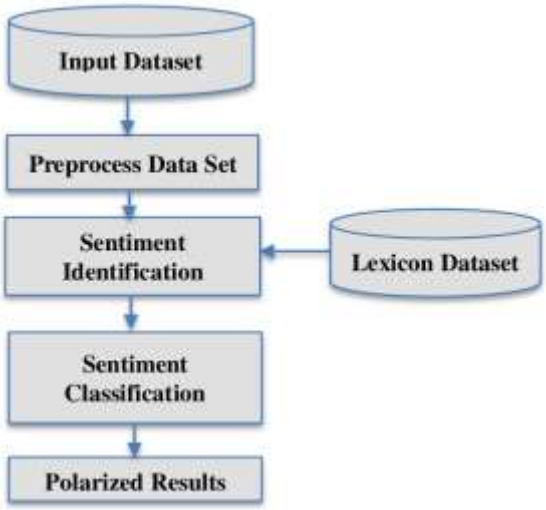

Fig 2: Lexicon-based sentiment analysis process

1. Input Dataset
The first step includes the retrieval of the data gathered from social media like Twitter and inserting it in analysis tool. For the analysis of sentiment, we need to clean the data which is retrieved to make the analysis easier. Cleaning the data means making the raw data in the format that the tool can understand. But while cleaning we do not need to remove the hashtags "#" as we are analysing it.

2. Pre-processing of Dataset
After successful input of dataset, the next step includes pre-processing of that data. The pre- processing technique is necessary because if we apply the analysis tool to raw data, it results in poor performance. The pre-processing involves following steps:

I. Tokenization
Tokenization is the step to divide the sentence into the set of word. In this step the word preceding with "#" is also marked as a different word. We divide the tweets from Twitter according to hashtags.

II. Stop word removal
Stop words are the words that add little meaning or no meaning in the analysis process. So, we need to remove it. The stop words include a, am, also, but, etc.

III. Stemming and Lemmatizing
Stemming and Lemmatizing are similar process. The only difference between them is, stemming trims the prefixes or suffixes of the word and lemmatizing makes use of more precise form i.e., dictionary to provide the root word. For example, the following table (considering the words preceded with "#")

Table 1: Stemming and lemmatizing difference

Word	Stemming	Lemmatizing
Dancing	Danc	Dance
Analysing	Analys	Analyse

3. Sentiment Identification
The sentiment of the word is identified on the basis of dictionary rules. The dictionary assigns value to each and every word as discussed earlier in this paper. So, this will decide the sentiment score of the data.

4. Sentiment Classification
In classification process, all the above steps and sentiment scores are combined together. If the statement includes two hashtag words, then the individual sentiments are calculated and then the difference between those two values gives the final sentiment. If one value is higher than other than the positive or negative is returned.

5. Polarized Results
Based on the dictionary, the system interprets whether the word is positive, negative or neutral.

Advantages and Challenges
[9] Lexicon-based approach has a wider term coverage, it almost does the complete sentiment analysis, but it is limited to finite number of words in lexicon and assigns the fixed sentiment score to words.
The corpus-based approach is able to generate a domain-sensitive, context-sensitive or topic-sensitive lexicon but the disadvantage is that a large corpus is required to capture the entire vocabulary of particular language. Whereas, dictionary based contain existing entire vocabulary of that language.

Discussion
On the basis of empirical study, we observed that Lexicon-based approaches utilize a lexicon to match chunks of the text with entity names. They also provide a nonlocal model for resolving multiple names matching the same entity resolution.
In [1], The data is collected by streaming API of twitter using python programming. They have done the analysis of hashtags of Greek language. They chose to examine hashtags appearing in 100 tweets. They removed a list of Greek stop-words from data, and replaced intonated characters with non-intonated ones, and changed every letter to uppercase. They used the proposed formula which is simply the arithmetic mean of each tweet and then used quadratic mean formula to decide the variance. Finally, they assigned maximum sentiment found in words contained in tweets, to the whole tweet. To analyse hashtag sentiment, they rejected most of the formula, as its results only depends on most sentimental tweet, not taking rest of the data into account. In [4], It discusses about the priority polarity scoring for prior polarization. It makes use of dictionary and extends it to WordNet to retrieve all the synonyms from it if that particular word is not found directly in the dictionary. And then it rates the word to find the sentiment.

Conclusion
Twitter is very demanding and highly used social media platform in recent times, which deals with a lot of data coming right away from the users in the form of tweets. The most popular aspect of Twitter is use of hashtags, which is used to express the feeling in the form of words preceding with hashtag symbol "#". This survey paper presents the proposed overview of Lexicon-based sentiment analysis and its process used for hashtags "#" analysis for the betterment of social media business. Lexicon-based analysis is the simplest way to analyse the sentiment. We have done the observational study about this method of analysing and discussed the advantages and challenges. We conclude that, some pre- processing steps such as stemming, lemmatizing, and replacing a word with its synonyms increase entity

matching rate. Lexicon-based approaches achieve better results for specific domains; however, they cannot identify new entities that are not in the lexicon, but it is most widely used approach for the sentiment analysis.

References

1. Sentiment analysis of Greek tweets and hashtags using a sentiment lexicon. Georgios Kalamatianos, Dimitrios Mallis, Symeon Symeonidis, Avi Arampatzis October2015 https://doi.org/10.1145/2801948.2802010
2. An analysis of 14 million tweets on hashtag-oriented spamming. Surendra Sedhai, Aixin Sun 28 May 2017 https://doi.org/10.1002/asi.23836
3. Sentiment Analysis on Twitter Data Varsha Sahayak, Vijaya Shete, Apashabi Pathan, International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 2015;1:2.
4. Sentiment Analysis of Twitter Data Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau Department of Computer Science, Columbia University, New York, NY 10027 USA.
5. Survey Paper on Sentiment Analysis: Techniques and Challenges Ansari Fatima, Anees Arsalaan, Shaikh Arbaz Shaikh, Sufiyan Shaikh. January 15, 2020.
6. An Overview of Lexicon-Based Approach for Sentiment Analysis A. Sadia, F. Khan, Fatima Bashir Published 2018.
7. Application and techniques of opinion mining Neha Gupta, Rashmi Agrawal, in Hybrid Computational Intelligence 2020. https://doi.org/10.1016/B978-0-12-818699-2.00001-9
8. Improved lexicon-based sentiment analysis for social media analytics Anna Jurek, Maurice D. Mulvenna & Yaxin Bi. Published 2015. https://doi.org/10.1186/s13388-015-0024- x
9. Corpus-Based Techniques for Sentiment Lexicon Generation: A Review October 2019 DOI:10.6025/jdim/2019/17/5/296- 305 Mohammad Darwich, Shahrul Azman Mohd Noah, Nazlia Omar, Nurul aida Osman