International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Twitter Sentiment Analysis

Riyaz. A. Jamadar[1], Priyadarshan. S. Chavhan[2]

[1]Assistant Professor, Department of Information Technology, AISSMS's Institute of Information Technology, Pune-411001, INDIA
[2]TE (Information Technology), AISSMS's Institute of Information Technology, Pune-411001, INDIA

ABSTRACT

Social media generates a large amount of sentiment-rich data in the form of tweets, status updates, blog posts, etc. In recent years, research on Twitter sentiment analysis has increased in a variety of fields. There are two approaches for sentiment analysis from text, namely the knowledge-based approach and the machine-learning approach. Nowadays, Many, researchers prefer to use machine learning algorithms for sentiment analysis. in which various machine algorithms like Decision Tree, Support Vector Machine and Naive Bayes are used to classify text into different sentiments.

1. Introduction to Twitter Sentiment Analysis

The Internet has changed the way people express their opinions about a product, service or website. Now this is happening through social media like twitter, facebook, instagram etc. when some want to buy some product online they will look for people's reviews on the product before making a decision. User-generated content is very large for the average user to parse. Various sentiment analysis techniques are used for automation. There are two basic approaches to sentiment analysis Symbolic or knowledge-based techniques and machine learning. A knowledge-based approach requires a very large database that has predefined emotions and an efficient knowledge representation for sentiment identification. A machine learning approach uses a training data set to develop a sentiment classifier that classifies sentiment. Since machine learning does not use a predefined data set, it is somewhat simpler than a knowledge-based approach. In this report, I will use different machine learning techniques to classify tweets Sentiment analysis is done at different levels, from coarse level to fine level. The coarse level deals with determining the sentiment of the whole document and the fine level deals with sentiment analysis at the attribute level. In the initial stages, twitter-specific features are extracted. Then these features are removed from the tweets to produce normal text. Then feature extraction is done again to get more features. This is the idea used in this paper to generate an efficient feature vector for Twitter sentiment analysis. Since there is no standard dataset for electronic device Twitter posts, we created a dataset by collecting tweets over a period of time. By performing a sentiment analysis on a specific domain, it is possible to identify the influence of domain information on the selection of a feature vector. Different classifiers are used to perform the classification to find out their influence in that particular domain with that particular feature vector.[1]

Motivation

1. An aspect of social media data such as Twitter messages is that they contain rich structured information about the individual involved in the communication. This can lead to more accurate semantic information extraction tools

2. It provides a means of empirically studying the properties of social networks. interaction.

1. Aim and Objective of the work

Aim

1). The aim of the project is to reveal people's sentiments based on the tweets they post about a product/service in order to understand consumer needs and improve the product/service accordingly.

Objectives

Implement an algorithm to automatically classify text as positive, negative or neutral.

Twitter sentiment analysis allows you to track what is being said about your product or service on social media and can help you uncover angry customers or previous negative mentions.

2. Literature Survey

2.1 **Twitter Sentiment Analysis Based on Ordinal Regression-** In recent years, research on Twitter sentiment analysis has grown rapidly to extract user sentiment about a topic, product, or service. Researchers prefer machine learning for such analysis. The paper focuses on detailed sentiment analysis of tweets based on ordinal regression using machine learning. The proposed approach consists of pre-processing tweets for an efficient feature using a feature extraction method and then scoring and balancing under different classes. Algorithms like SVM,DT,RF are used for sentiment classification after feature extraction and balancing text score.

2.2 **Recognizing Contextual Polarity in Phrase-Level Sentiment-** This paper presents a new phrase-level sentiment analysis approach that first determines whether an expression is neutral or polar and then clarifies the polarity of polar expressions. With this approach, the system is able to automatically identify the context polarity for a large subset of sentiment expressions and achieve results that are significantly better than the baseline.

2.3 **Robust sentiment detection on Twitter from biased and noisy data-** The paper proposes an approach for automatic sentiment detection in messages (tweets) on Twitter that examines some characteristics of how tweets are written and meta-information about the words that make up these messages. labels were provided by several sentiment detection websites over Twitter data.

2.4 **Opinion Mining on Social Media Data-** Since there is a lot of raw data about people who post real-time messages about their opinions on various topics of daily life, it is worth trying to collect and analyze this data, which can be useful for users or managers to get information . decision. this paper proposes a new system architecture that can automatically analyze the sentiment of these messages. the system manually annotates data from Twitter, one of the most popular microblogging platforms, for sentiment analysis. In this system, machines can learn how to automatically extract a set of messages that contain opinions, filter out non-opinion messages, and determine the direction of their sentiment.

3. A brief Intro to Twitter Sentiment Analysis

The first step in sentiment analysis is the collection and sorting of data. There is a sea of data on Twitter, it is important to pick the data that is most relevant to the problem you are looking to solve or the thing you wish to find out. Only relevant data can be used to train the sentiment analysis model and test whether the model performs satisfactorily on Twitter data. Another important aspect to cover is what type of tweets you are looking to analyze – historical or current. To sort this data, you first need to extract it from Twitter.

Once the data is gathered and sorted, it then needs to be cleaned before it can be used to train the Twitter sentiment analysis model. Twitter data is mostly unstructured, so the cleaning process involves removing emojis, special characters, and unnecessary blank spaces. The process also includes getting rid of duplicate tweets, making format adjustments, and also removing very short tweets – those less than three characters. Clean data can give more accurate results.

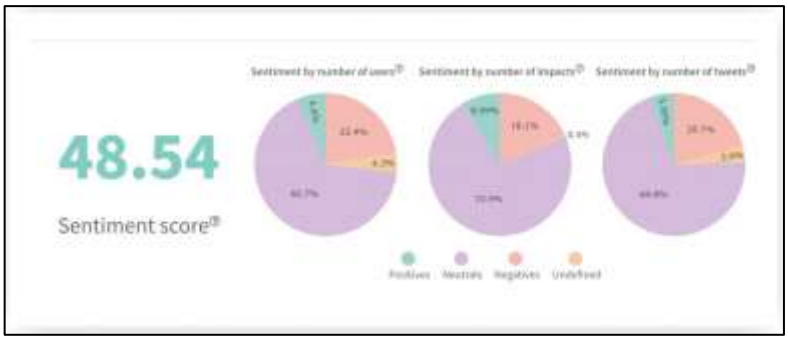

Fig. 1. Twitter sentiments Analysis using with Twitter Binder

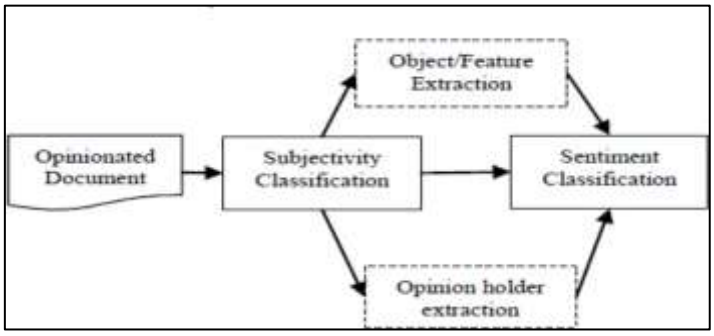

Fig. 2 Sentiment Analysis Tasks

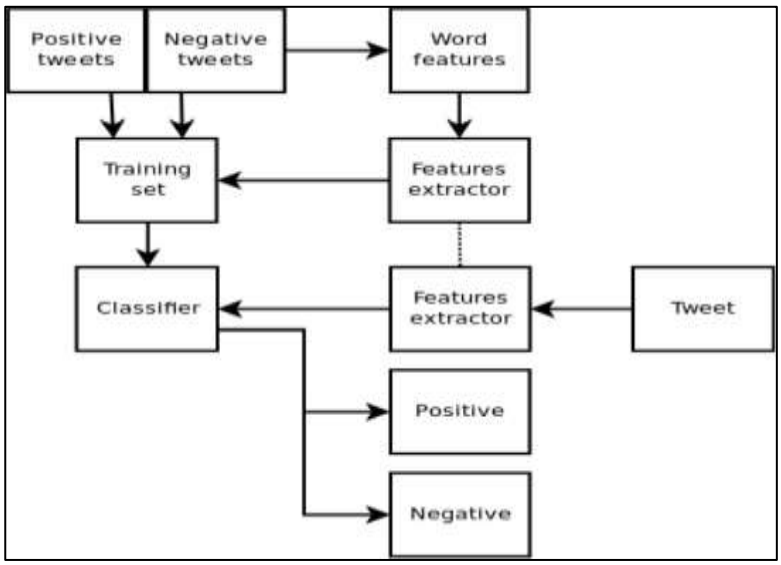

Fig. 3 Sentiment Analysis Architecture

4. Methodology:

Tweets are short messages full of slang words and typos. So we perform sentiment analysis at the sentence level. This is done in three stages. In the initial stage, pre-processing is performed. An element vector is then created using the appropriate elements. Finally, using different classifiers, the tweets are classified into positive and negative classes. on the basis number of tweets in each class, the final sentiment is derived.

1. Creation of Dataset:-

Collecting tweets automatically using Twitter API. and manually annotating them as positive, negative, or neutral.

2. Pre-processing of tweets:-

Keyword extraction is difficult on Twitter due to typos and slang words. To avoid this, a preprocessing step is performed before extracting the feature. Preprocessing steps include

• Removal of all hyperlinks ("http://url"), hashtags ("hashtag"), retweets (RTs) and links to usernames ("@username") that appear in Tweets;

• Removal of repeated letters

• Avoid typos and slang words

• Converting words to lowercase letters

• Removal of commonly used words that have no special meaning, e.g

such as pronouns, prepositions and conjunctions.

• Reduction of words to their stem or common root by removing plurals, genders

• Segmenting sentences into words or phrases called tokens by discarding some characters

• Removal of duplicate data tweets

• Replacing all emoticons with their corresponding sentiment.

3. Creation of Feature Vector:-

Feature extraction is performed in two steps. In the first step, Twitter-specific features are extracted. Hashtags and emoticons are relevant specific features of Twitter. Emojis can be positive or negative. So they have different weight. Positive emoticons have a weight of "1" and negative emoticons have a weight of "-1". There can be positive and negative hashtags. Therefore, the number of positive hashtags and negative hashtags are added as two separate features in the feature vector. Twitter-specific features may not be present in all of the tweets. Therefore, additional feature extraction needs to be performed to obtain additional features. After the twitter-specific features are extracted, they are removed from the tweets. Tweets can then be considered as plain text. Then, using a unigram approach, tweets are represented as a collection of words. We maintain a list of negative keywords, a list of positive

keywords, and a list of various words that represent negation. The counts of positive and negative keywords in the tweets are used as two different functions in the feature vector. The presence of negation adds a lot to the sentiment. So their presence is also added as a relevant feature. In the case of multiple positive and negative keywords, not all keywords can be treated equally. Therefore, a special keyword is selected from all tweets. For tweets that contain only positive keywords or only negative keywords, a search is performed to identify a keyword that has relevant parts of speech. A relevant part of speech is an adjective, adverb or verb. Such relevant part of speech is defined based on their relevance in determining sentiment. Keywords that are adjectives, adverbs, or verbs show more emotion than others. If a relevant part of speech can be determined for a keyword, it is considered a special keyword. Otherwise, the keyword is randomly selected from the available keywords as a special keyword. If there are both positive and negative keywords in the tweet, we will select any keyword that has relevant parts of speech. If the relevant part of speech is present for both positive and negative keywords, neither is selected. A keyword special property has a weight of "1" if positive, "-1" if negative, and "0" if not. The part-of-speech feature has a value of '1' if relevant and '0' otherwise. So the feature vector consists of 8 relevant features. The 8 features used are part of speech tag, special keyword, presence of negation, emoticon, number of negative keywords, number of positive keywords, number of positive hash tags, and number of negative hash tags.

4. Sentiment Classification:-

Nave Bayes Classifier The Nave Bayes classifier takes all the features in the feature vector and analyzes them individually because they are equally independent of each other. The conditional probability for Naïve Bayes can be defined as P(X—yj) = label classes. Here in our work, there are various independent features such as emoticons, emotional keyword, number of positive and negative keywords, and number of positive and negative hash tags, which the Naive Bayes classifier effectively uses for classification. Nave Bayes does not consider relationships between elements. Thus, it cannot use relationships between part of speech, emotional keyword, and negation.[2]

5. Advantages and Disadvantages

Advantages:-

The use of these facts can be implemented to make wiser choices associated with the use of resources, to make improvements in organizations, to offer better products/services and ultimately to improve the lifestyle of citizens.

Disadvantages:-

Analyzing tweets is an example of this, as they are usually associated with hash tags, emoticons and links, making it difficult to determine the sentiment expressed. On top of that, there is a need for automatic techniques that require large datasets of annotated posts or lexical databases where emotional words are associated with a sentiment value.

6. Applications of Twitter Sentiment Analysis

There are different application of Twitter Sentiments Analysis as follows:

1) Social Media Monitoring:- Twitter sentiment analysis allows you to keep track of what's being said about your product or service on social media, and can help you detect angry customers or negative mentions before they they escalate.

2) Customer service:- Twitter sentiment analysis allows you to track and analyze all the interactions between your brand and your customers, so you can make sure you respond to the most critical issues first.

3) Marker research:- You can use Twitter sentiment analysis to track specific keywords and topics to detect customer trends and interests

4) Political Campaigns:- A huge part of Twitter conversation revolves around news and politics. During the US 2016 elections, we performed Twitter sentiment analysis using MonkeyLearn to analyze the polarity of Twitter mentions related to Donald Trump and Hillary Clinton.

7. Conclusion

Machine learning and knowledge-based techniques are used to identify sentiment from text. Machine learning techniques are simpler and more efficient than knowledge-based techniques. There are some challenges in identifying an emotional keyword from tweets with multiple keywords. It is also difficult to manage misspellings and slang words. To solve these problems, feature extraction is performed, which is performed in two steps to create an effective feature vector. In the initial step, Twitter-specific features are extracted and added to the feature vector. Then these features are removed from the tweets and feature extraction is performed again as if it were normal text. The classification accuracy of the feature vector is tested using various classifiers such as Nave Bayes, SVM, Maximum Entropy and Ensemble classifiers. This feature vector works well for the domain of electronic products.

8. References:-

[1] Evolutionary Machine Learning Techniques: Algorithms and Applications (Algorithms for Intelligent Systems) by Seyedali Mirjalili, Hossam Faris, Ibrahim Alijarah.

[2] Sentiment Analysis using Maximum Entropy Algorithm in Big Data Durgesh Patel , Sakshi Saxena , Toran Verma, International Journal of Innovative Research in Science, Engineering and Technology(http://www.ijirset.com/upload/2016/may/246_49_Sentim ent.pdf)

[3] International Research Journal of Engineering and Technology (IRJET) Volume: 03 Issue: 08 | Aug-2016 www.irjet.net, Sentiment Analysis Using SVM and Maximum Entropy Snehal L. Rathod, Sachin N.Deshmukh

[4] International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-8 June, 2019, Sentiment Analysis Using Naïve Bayes Classifier Sentiment Analysis Using Naïve Bayes Classifier