International Journal For Technological Research In Engineering
Volume 10, Issue 9, May-2023 *ISSN (Online): 2347 - 4718*

SENTIMENT ANALYSIS OF TWITTER POSTS USING MACHINE LEARNING APPROACHES - A REVIEW

[1]Amit Kumar, [2]Prof. Ritesh kumar
[1]Research Scholar, [2]Assistant Professor
DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
PKG Group of Institutions
KURUKSHETRA UNIVERSITY, KURUKSHETRA

Abstract— This study deals with the issue of sentiment analysis on Twitter, which involves categorising tweets based on whether they exhibit positive or negative emotions. As all know, 'Twitter' is a platform that connects news and social networking issues to online social media micro blogging website that enables users to post 140-character limit to an issue update. It has more than 200 million users as a registered end user [24], of whom 100 million are active users and half of them log in at least once in a day, it is a free online service that is growing in a strictly increasing manner. Each day, it generates about 250 million tweets [20]. In this work, we represent a public opinion by analyzing sentiments stated in the tweets in light of this significant use. Numerous applications need to justify public mood, including businesses environment and deals to gauge the market reaction to their goods, the prediction of political outcomes, and the analysis of socioeconomic phenomena like the stock exchange, etc. The goal of this project is to create a practical classifier that can accurately and automatically identify the sentiment of an unidentified Twitter stream.

Keywords: Machine Learning, Twitter

1. INTRODUCTION

This Twitter sentiment analysis project falls under the categories of "Pattern Classification" and "Data Mining." Both of these concepts are intimately connected and entangled, and they may be properly described as the automated (unsupervised) or semiautomatic process of catching "useful" arrangements in vast sets of data (supervised). In order to effectively identify individual unlabeled data samples (tweets) which come out to be the best version of the pattern model with the best characteristics accordingly, the project will mainly depend on "Natural Language Processing" methods for withdrawing considerable patterns and features from the enormous dataset of tweets.

The separation of formal language-based appearance and informal blogging-based characteristics may be characterized into two broad groupings that can be utilized for modeling patterns and categorization. Language-based characteristics deal with formal linguistics that carries the parts of speech in which each sentence is specified with the preceding sentiment polarity of certain words and phrases. Prior sentiment polarity distinguishes the inherent inbuilt inclination of certain words and phrases to communicate particular and specific feelings in general. For instance, the term "great" has a strong connotation of positivity, while the word "evil" carries a strong connotation of negativity. Therefore, anytime a word with a meaningful sense is employed in a phrase, there is a great chance that the statement as a whole will be fruitful. On the other hand, a syntactical approach to the issue is Parts of Speech labeling. It indicates to the ability to recognize the part of speech in a phrase corresponding to each word, including a noun, pronoun, adverb, adjective, verb, interjection, etc., belongs to. By examining the frequency distribution of these parts of speech in a particular category of labeled tweets, either alone or in the composition of other parts of speech, patterns may be derived. The elements of Twitter are more infrequent, attach to how individuals express themselves on social networking online platforms, and compress their ideas into the constraint of 140 characters that Twitter provides. Twitter hashtags, retweets, word capitalization, word lengthening, question marks, the use of URLs in tweets, exclamation points, online emoticons, and internet shorthand/slang are a few examples of these. The two kinds of classification methods are supervised vs. unsupervised, and non-adaptive vs. adaptive/reinforcement methods. When we have pre-labeled data samples available, we may train our classifier using an instructed technique.

2. LITERATURE REVIEW

Limitations of Prior Art

Since sentiment analysis in the context of microblogging is still a relatively unexplored field of study, there is still much to be discovered. There has been a substantial amount of similar past work on phrase-level sentiment analysis as well as sentiment analysis of user reviews [x], papers, online blogs/articles, and documents. These are distinct from Twitter mostly because of the 140-character character restriction per tweet, which pushes users to deliberate opinions in a concise manner. The classification of the best sentiment outcomes is determined by incorporating the instructed learning methods such as Bernoulli Naive Bayes, however, the human labeling needed for this method is quite costly. Unsupervised and semi-supervised techniques have received some attention, but there is still much space for improvement. Many researchers experimenting with novel characteristics and classification methods simply compare their findings to baseline performance. In order to choose the best features and most effective classification approaches for specific applications,

appropriate and formal comparisons between these findings obtained by various features and classification techniques are required.

Related Work

Due to its simplicity and effective performance, the feature model named 'the bag-of-words model' used in a frequent manner has practical application in all text classification problems. The approach considers the text to be categorized as a bag or bundle of distinct words with no connection between or dependent on one another; hence, it entirely ignores the grammar and word order within the text. This model has been used by several researchers and is also particularly well-liked in the sentiment analysis field. Using unigrams as features is the easiest technique to fix the model into the classifier. In our literature, an n-gram is often defined as a continuous series of "n" words that stand alone from all other words and grammatical structures. Unigrams are just a grouping of distinct words in the text that need to be categorized, and we make the assumption that the presence or absence of other words in the text will not have an impact on the likelihood of recurrence of any given word. Although it is a relatively simplistic assumption, it has been shown to provide rather acceptable performance. Assigning unigrams polarity under a certain priority and averaging the overall polarity of the text are two straightforward methods for using unigrams as features. The overall polarity of the text may be computed by adding the previous polarities of individual unigrams. If a term is often used to indicate something good, then its prior polarity would be positive.

If a term is often linked with positivity, like "sweet," it would be positive; if it is normally connected with negativity, like "evil," it would be negative. Degrees of polarity, or how suggestive a word is for a certain class, may also be included in the model. The word "amazing" would likely have high subjective polarity in addition to being positive, but the term "decent" would likely have weak subjectivity despite having positive previous polarity.

The prior polarity of words may be used as a feature in three different ways. Using publicly accessible internet lexicons or dictionaries that map a word to its preceding polarity is the easier unsupervised method. An online tool called the Multi-Perspective-Question-Answering (MPQA) has a subjectivity lexicon that categorizes 4,850 terms as either "positive" or "negative" and as having "strong" or "weak" subjectivity. Another tool that indicates the likelihood that a word is positive or negative is SentiWordNet 3.0. The second method involves building a unique prior polarity dictionary from our training data based on how often one word appears corresponding to that specific class. To give an instance, if the appearance of a certain more is more often as a positive labeled phrase in our training dataset (as opposed to other classes), we may determine that the likelihood that the word belongs to the positive class is greater than the likelihood that it will appear in any other class. It has been shown that the proposed strategy is better in performance since the prior polarity of the words is quite similar and fitted easily into a certain sort of text. It is claimed that this strategy is not as generic as in the previous approach.

The latter, however, requires supervision since the training data must first be classified into the proper classes in order to determine the relative frequency of a word in each class. When the lexical word features were combined with the parameters/features of unique n-gram words created from the training data, Kouloumpis et al. found that performance was worse than when the n-grams were taken in a single count.

3. MOTIVATION

As many traditional online articles and web blogs are available in the literature, in a contrast, we believe that twitter is an online social -media platform that provides us a more accurate representation of the opinions on a current topic. The logic behind is that, if we compare to conventional blogging platforms, twitter has a considerably higher volume of data that is quite relevant to the current material. On an addition, the reactions/opinions on Twitter are very quick to mark and broader (since the counting of the users of Tweeter is high as compare to those users who write the blogs on web pages on a regular manner). In macro-scale socioeconomic situations like stock markets where we anticipated the rate of a certain company/firm, the research is very crucial and complex about the public sentiment/approaches. This might be accomplished by examining the general opinion of the company/firm over time and utilising several economics methodologies to establish the relationship between the opinions of the public and the market evaluation of the firm's stock. Since Twitter enables us to collect (download) several streams of geo-tagged tweets for many specified areas as business developers may also predict how reliable their product is and how it is going to react in the market and the consumers and which market having a favourable reaction or unfavourable reaction. If system developers can estimate this data, they may determine directly the causes of regional based on several feedback and sell their items (products) at their desirable cost in order to develop their business by implementing the relevant solutions like forming appropriate market groups, etc. Another benefit of developing the sentiment analysis is to know the system behaviour so that we can predict the outcomes and the several measures of popular political elections and surveys. In an effective research article, which was analysed in Germany for the purpose of forecasting the outcomes of federal elections, Tumasjan et al. came to the conclusion that Twitter is a good indicator of offline mood.

REFERENCES

[1] Albert Biffet and Eibe Frank. Sentiment Knowledge Discovery in Twitter Streaming Data. Discovery Science, Lectures Notes in Computer Science, 2010, Volume 6332/2010, 1-15, DOI: 10.1007/978-3-642-16184-1_1

[2] Alec Go, Richa Bhayani and Lei Huang. Twitter Sentiment Classification using Distant Supervision. Project Technical Report, Stanford University, 2009.

[3] Alexander Pak and Patrick Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In Proceeding of international conference on language Resources and Evaluation(LREC), 2010.

[4] Andranik Tumasjan, Timm O. Sprenger, Philipp G.

Sandner and Isabell M. Welpe. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In Proceeding of AAAI conference on Weblogs and Social Media (ICWSM),2010.

[5] Google App Engine https://developers.google.com/appengine/.

[6] Google Chart API <https://developers.google.com/chart/>

[7] Tweet Stream: Simple Twitter Streaming API Access http://pypi.python.org/pypi/tweetstream

[8] Twitter REST API <https://dev.twitter.com/docs/api>

[9] Twitter Sentiment, an online application performing sentiment classification of Twitter. <http://twittersentiment.appspot.com/>

[10] Ian H. Witten, Eibe Frank & Mark A. Hall. Data Mining – Practical Machine Learning Tools and Techniques.