Sentiment Analysis of Tweets

Rakhin Mostafa, Annajiat Alim Rasel, Md. Mustakin Alam, Humaion Kabir Mehedi

Introduction

Every day, millions of people across the world utilize social networking platforms to communicate with one another and share their views, interests, and personal information [1] However, people do write about a wide range of subjects, from social gatherings to product evaluations. Users may educate and persuade others by participating in online communities' interactive forums. Social media also provides companies with a forum for communicating with their target demographic, whether via paid advertising or direct conversation about the company's products and services. On the other hand, viewers may decide for themselves what they want to see and how they react. The good and bad news about the business may spread rapidly this way. However, consumer behavior and choices are susceptible to the persuasion of social networks. To provide just one example, [2] claims that 87 percent of online shoppers rely on reviews before making a final choice. Because of this, it is important for a business to improve its ability to understand its customers' perspectives in order to better anticipate their needs and develop a competitive strategy.

Objective

The major purpose of this research is to learn how to use sentiment analysis in microblogging to examine user reactions to a company's wares. The secondary objective is to design a system for collecting and organizing user feedback on goods in a way that makes it easy to do sentiment analysis on a huge volume of tweets.

Twitter

Twitter is a well-liked microblogging service where users may instantly communicate with one another using short messages (called "tweets") of no more than 140 characters in length [3]. Users' tweets cover a broad variety of issues they encounter in their daily lives. For quick and easy polling of public opinion, Twitter is a fantastic resource [9]. A corpus of tweets is essential for sentiment analysis, also known as opinion mining and natural language processing [1]. Social media platform Twitter has over 500 million users and facilitates the daily exchange of millions of tweets, making it a useful resource for businesses interested in monitoring consumer sentiment toward their own and rival brands and goods [10]. Tweets, reviews, blogs, and discussions in online forums are all examples of social media-generated opinions that can be analyzed in [2], demonstrating that the internet is the most efficient, extensive, and easily accessible medium for sentiment analysis. The length of postings is the primary distinction between a standard blogging platform and a microblogging one like Twitter. Twitter's character count limits are there to promote rapid communication by encouraging users to get to the point. Only recently have both small businesses and multinational firms begun to fully appreciate the potential of microblogging as a tool for promoting e-commerce [3]. However, a few years ago [3, 4] a foreign microblogging network called Twitter was formed for promoting international trade websites.

It is clear that the instant of sharing, interactive, and community-oriented features is launching a new bright spot in the e-commerce industry and allowing businesses to interact with brand marketing strategy, product positioning, increased product sales, and talk to customers for positive interaction[8]. According to [9], manufacturers have started monitoring users' reactions to their goods on social media. These companies routinely track user comments made on microblogs and post replies to those comments [11].

Social Media

Social media" refers to web-based programs that make it easier to create and share user-generated content [12]. According to a debate by Internet World Start [14], in 2012, Americans spent a total of 121 billion minutes on mobile devices and social media, up from 88 billion minutes in 2011. This pattern indicates that people are spending more and more time online, particularly on social networking sites. Despite the fact that businesses rely on social networking sites to stay in touch with their clientele, studies have shown that such activities have a deleterious effect on workplace efficiency [12]. Because of the ease with which material may be shared on social media platforms, sensitive information may be leaked [11]. However, [13] argued that there are benefits to social media beyond socializing, including boosting brand awareness and revenue. Online social learning, employee searches, corporate marketing, and candidate recruitment were highlighted by [15]. The act of buying and selling products and services online, typically through social media sites like Twitter, is known as e-commerce. This has many advantages, including the ability to transact business at any time of day or night, a reduced need for in-person interactions, and a global reach. A growing number of companies are using social media to gather data on industry trends, consumer habits, and brand views.

Problem affirmation

Even if there are tools to extract information about a person's opinions about a given product or service, organizations and other data workers still come into challenges with data extraction.

-Web-based application users' feelings analyzed Pay attention to only one tweet at a time.

With the proliferation of social media sites like Twitter, there are now potentially enormous volumes of opinion texts available for sentiment analysis [3]. For a person, this would translate to an overwhelming volume of information, making it challenging to quickly extract terms, read them, assess them, and summarize/organize them in a way that is understandable.

-The difficulty of employing sentiment analysis for terrible English.

The use of slang and colloquialisms in everyday discourse is an example of informal language. Some systems may struggle to understand informal language, which might cause problems in analysis and decision-making. Shorten or spoken languages like ‘ Should not ‘ and ‘shouldn't’.

Emoticons, which are graphical representations of human facial emotions [5], allow a recipient to better comprehend the speaker's intended tone or mood in a purely spoken exchange [6]. A common example would be the sign for joy. The present algorithms do not have enough information to derive emotions from emoticons. When words fail, people turn to emoticons [6] to convey their feelings and thoughts. If the company can't figure this out, it's in a bind. Even when using SMS or other short messaging services, the abbreviated version is often preferred (SMS). Twitter users will increasingly resort to abbreviations in an effort to limit their posts to 140 characters. This is because the maximum number of characters in a tweet is 140 [7].

Putting the datasets through some preliminary processing

The most prominent A tweet contains the varied and often conflicting opinions of its author(s). The Twitter dataset that was used for this investigation has previously been split into two groups: those with a negative polarity and those with a positive polarity, making it straightforward to do sentiment analysis on the data and evaluate how different variables affect sentiment. Since polarity is present in the raw data, it is more likely to have redundancies and inconsistencies. All of the following are part of the preprocessing of tweets:

-Remove any references to websites (www.xyz.com), hashtags (#xyz), or users (@username) in your message.

-Spell check and cope with sequences of characters that appear more than once.

-Replace all the emoticons with your own interpretation of their meaning.

-Get rid of the decimal points, commas, and dashes.

-Deepen Acronyms and Drop Fillers (we can use an acronym dictionary)

-Tweets published in a language outside English should be deleted.

Data-Rich Extras

Twitter allows its users to send and read "tweets," or short, time-sensitive communications, in real-time. Twitter's unique format and qualities provide additional hurdles for sentiment analysis and shape the way this research is undertaken on the platform.

Some fundamental features of tweets are as follows:

Message length:- You're limited to 140 characters for your tweet. This differs from prior research on sentiment classification, which primarily concerned itself with classifying lengthier texts like product and film reviews.

Style: Twitter posts are more likely to include typos and internet lingo than those in other fields. People write short, concise communications full of misspellings, emojis, and acronyms.

Accessibility of Data

We have access to a wealth of data. More users tweet in the open domain, making data more accessible, compared to Facebook's extensive privacy options. The Twitter API facilitates the collection of tweets for instructional purposes.

Twitter users send messages on a wide range of subjects, in contrast to topic-specific sites. Unlike a lot of the older studies, which focused on narrow topics like movie reviews, these new ones aim to cover a wider range of topics.

Actual time

Because of their inherently lengthy and labor-consuming nature, blogs are not updated in real-time. Tweets, on the other hand, are limited to 140 characters and are constantly being revised. This catches the first reactions to events and gives them a more immediate sense.

Now we'll go over some Twitter basics:

Emoticons:

These phrases are written out using just letters and punctuation. It is common practice for users to express themselves via the usage of emoticons.

Target:

Users of Twitter address one another using the "@" symbol. Everyone who has been referenced in this manner receives an alert.

Identifiers using a hash function:

The "#" symbol is used to indicate a subject for discussion. Twitter users use it to raise the profile of their messages.

Unique signs:

The "RT" symbol indicates that the tweet is a repost from another user.

Obtaining Features

A great deal of novelty has been preserved in the preprocessed dataset. We are able to extract the features from the revised dataset by making use of a method that is known as "feature extraction." After that, this characteristic is applied to determine the positive and negative polarity of the device of a phrase using models such as the unigram and the bigram, which are useful for gauging people's views [15].

Processing text or documents requires machine learning methods that can accurately reflect their salient characteristics. During categorization, these salient features, or "feature vectors," are used. Among the traits that have been discussed in the academic literature are:

Words and phrases:- Unigram, bigram, and n-gram counts,are among the available features. Studies have shown that word presence, rather than word frequency, is a more accurate way to characterize this trait. The findings improved when presence was used instead of frequency.

Labels for the linguistic structure:- the Organization of Language Effective indicators of subjectivity and emotion include adjectives, adverbs, and certain classes of verbs and nouns. Syntactic dependence patterns are generated through parsing dependency trees.

Opinionated Phrases & Words:- Features may include not just individual words, but also idioms and expressions of emotion. a man an arm and a leg, to provide just one example.

conditions in the position:- How much a particular phrase alters the overall tone of a text depends on where in the text it is used.

Denial:- The intricacy of the negative obscures its meaning. Adding a negative may flip the polarity of a statement.

Grade Grammar:- Syntactic patterns, including collocations, are often used as characteristics in studies on learning subjective patterns.

An Examination of Twitter Sentiment

The tone of a tweet or remark may be used as an indication of public opinion for a number of reasons [11]. Words having positive and negative connotations were also mentioned as ways in which feelings may be broken down. To determine how people are feeling or what they think about a certain topic, researchers use a natural language processing technique called sentiment analysis [8]. Semantic orientation, or the polarity and intensity of words within a text, may be separated from the text's subjective tone by a method called sentiment analysis [12]. The lexicon-based and machine-learning-based approaches are employed most regularly[11].

First, A Lexicon-Based Method

A lexicon-based method relies on a set of words that have already been associated with feelings. The lexicon methods determine the document's orientation depending on the papers' semantic orientation, however, this is conditional on the context in which the lexicon methods were first developed [8]. Furthermore, a lexicon sentiment's goal is to identify words in the corpus that transmit opinion and then anticipate how that opinion will be represented in the text, as stated in [4]. The lexical techniques have been shown [9], and they all adhere to the same basic paradigm:

1. Before tweeting, please remove all punctuation.

2. Make sure that (s), the polarity score, is equal to zero.

3. Checking if the token is present in a dictionary yields a positive (+) S value if the token is positive and a negative (-) S value if the token is negative (-).

4. Take a look at the tweet's polarity score as a whole. Tweet something encouraging if s is greater than the cutoff.

Nonetheless, as mentioned in [14], one benefit of the learning-based approach is that it may be used to produce trained models that are customized to a certain application or circumstance. On the other hand, the availability of data with appropriate labels and the following constrained use of the new data strategy [14] may make labeling data expensive or even troublesome for certain operations.

Second, A Method based on Machine Learning

In supervised classification methods, often used by machine learning algorithms, sentiment analysis is typically presented as positive or negative [4]. This approach requires labeled data for classifier training [5]. It is evident that the negative (e.g., Not handsome) and positive (e.g., Very Handsome) local contexts of a term need to be taken into account when using this approach [6]. However, as shown in [4], a basic paradigm for constructing a feature vector may be summarized as follows:

1) Label each tweet with a section of the address.

2) make a list of all the adverbs that have been used in your tweets.

3) List the top N adjectives.

Navigate the experimental tweets to produce:
Positive and negative word counts, plus word frequency
beautiful (+5) became not beautiful (switch negation) (-5). Example:
She's not great (8-3=5) or terrible (-8+3=-5). The shift in value reflects the negation of a strongly positive or negative value in this case. [21] says a machine-learning-based approach is better for Twitter than a lexical-based approach. Machine learning may output a preset number of popular terms, with each word's Twitter frequency as an integer.

Strategies for analyzing sentiment

Entity semantics collected based on tweets could be used to evaluate the degree to which a set of entities is associated with a certain polarization [12]. Polarity, in its most basic sense, refers to whether a paragraph or statement is optimistic or pessimistic [5]. However, methods from sentiment analysis, such as, may be used to attribute polarity like:-

1. **Processing of Natural language (NLP)**

The cornerstone of natural language processing (NLP) methods is statistical learning [6],The program utilizes a corpus to learn the rules. Sentiment analysis is one facet of NLP that has been used at varying levels of detail. A document-level [7], then a sentence-level [8], and finally a phrase-level [13] approach have all been taken so far. Natural Language Processing (NLP) is a subfield of computer science that focuses on making computers understand human language and input.

2. Machine Learning with a Support Vector Regression (SVM):
The purpose of a Support Vector Machine (SVM) for emotion detection is to tweet. In addition to what the SVM has already been able to do, Use data mining and extraction techniques to improve accuracy to between 70% and 81.3 data from a test. The SVM was trained on these imperfectly labeled data. The team had an 81.3% success rate in classifying emotions.

Conclusion

The growing popularity of microblogging sites like Twitter has opened up a new window of opportunity for the study and use of methods for detecting and mining user sentiment. Some innovative approach to analyzing Twitter data for sentiment is given in this research.
In this study, we give a complete evaluation and in-depth comparison of the current state-of-the-art procedures for opinion mining. Some of these methodologies include machine learning and lexicon-based methods, as well as methods that work across domains and languages and evaluation metrics. Machine learning techniques such as support vector machines (SVM), which have been the subject of research, have been shown to have the highest accuracy and may be considered baseline learning techniques. On the other hand, lexicon-based techniques are extremely successful in certain circumstances and require very little effort when processing documents that have been human-labeled. We also analyzed how adding or removing characteristics affected the classifier. Our research led us to the conclusion that the more precise the data, the more precise the findings. When compared to other models, sentiment accuracy is improved when a bigram model is used. To enhance sentiment classification accuracy and adaptive capability across domains and languages, we may investigate merging machine learning with the opinion lexicon approach.

References

[1] A.Pak and P. Paroubek. „Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010, pp.1320-1326

[2] . Sharma, “Application of Support Vector Machines for Damage detection in Structure,” Journal of Machine Learning Research, 2008

[3] Bifet and E. Frank, "Sentiment Knowledge Discovery in Twitter Streaming Data", In Proceedings of the 13th InternationalConference on Discovery Science, Berlin, Germany: Springer,2010, pp. 1-15

[4] Socher, Richard, et al. "Recursive deep models for semantic compositionality over a sentiment Treebank."Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP). 2013.

[5] B. Pang, and L. Lee, “Opinion mining and sentiment analysis,” 2nd workshop on making sense of Microposts. Ithaca: Cornell University. Vol.2(1), 2008

[6] J. Kamps, M. Marx, R. J. Mokken, and M. De Rijke, “Using wordnet to measure semantic orientations of adjectives,” 2004

[7] L. Colazzo, A. Molinari and N. Villa. “Collaboration vs. Participation: the Role of Virtual Communities in a Web 2.0 world”, International Conference on Education Technology and Computer, 2009, pp.321-325

[8] L. Barbosa, J. Feng. “Robust Sentiment Detection on Twitter from Biased and Noisy Data”. COLING 2010: Poster Volume, pp. 36-44.

[9] A. Bifet and E. Frank, ”Sentiment Knowledge Discovery in Twitter Streaming Data”, In Proceedings of the 13th International Conference on Discovery Science, Berlin, Germany: Springer,2010, pp. 1–15.

[10] A. Kumar and T. M. Sebastian, “Machine learning assisted Sentiment Analysis”. Proceedings of International Conference on Computer Science & Engineering (ICCSE’2012), 2012, pp. 123-130

[11] M.Rambocas, and J. Gama, “MarketingResearch: TheRoleof SentimentAnalysis”. The 5th SNA-KDD Workshop’11. University of Porto, 2013

[12] M. Comesaña, A. P.Soares, M.Perea, A.P. Piñeiro, I. Fraga, and A. Pinheiro, “ Author ’ s personal copy Computers in Human Behavior ERP correlates of masked affective priming with emoticons,” Computers in Human Behavior, 29, 588–595, 2013

[13] A. K. Jose, N. Bhatia, and S. Krishna, “TwitterSentimentAnalysis”. NationalInstituteof TechnologyCalicut,2010.

[14] S. Sharma, “Application of Support Vector Machines for Damage detection in Structure,” Journal of Machine Learning Research, 2008

[15] B. Pang, and L. Lee, “Opinion mining and sentiment analysis,” 2nd workshop on making sense of Microposts. Ithaca: Cornell University. Vol.2(1), 2008.