Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/4/20230354

Sentiment analysis for social media using SVM classifier of machine learning

Qingyu Huang

College of Liberal Arts & Sciences, University of Illinois at Urbana-Champaign, Champaign, Illinois, 61820, United States

qingyuh2@illinois.edu

Abstract. The community's perspectives and comments are a valuable resource for businesses and other organizations. In the past, businesses used inefficient procedures. Now that social media is the new trend, it enables an unprecedented level of analysis and evaluation. This enables unprecedented analysis and evaluation of various factors. This enables unprecedented analysis and evaluation of a wide range of topics and components in different contexts and settings. Throughout business history, these strategies have been expected. This field of study is called "sentiment analysis." SVM was used to analyze sentiment for this research project. One of these duties required an SVM(SVM). Support vector machines, or SVM, is a popular supervised machine learning algorithm for determining text polarity. SVM abbreviates support vector machines. Precision, recall, and F-measure are used to evaluate SVM using two datasets of pre-classified tweets. Tables and graphs are used to communicate research findings. This research classifies tweets about US-Airlines and performs sentiment analysis with an accuracy of 91.8 percent, precision of 91.3 percent, and recall of 82.3 percent, as well as the F1 of 86.9 percent.

Keywords: SVM, Classifier, Sentiment Analysis, Emotions, US-Airlines.

1. Introduction

It is anticipated that over the next few years, there will be an increase in demand for text mining tools and methods that are both effective and efficient. This demand is being driven by the massive amounts of textual data currently present in the modern world. Because of online social networking websites, this information keeps growing and expanding daily (Facebook and Twitter). Mining the opinions and stances expressed in such massive amounts of data and reviews can provide organizations with innumerable benefits. Businesses can efficiently maintain and improve their market position by utilizing sentiment analysis. These steps include figuring out which products or services need to be improved, which price allocations most people are unhappy with, and which new features the community wants. Sentiment analysis enables these businesses to do so. They can take effective and efficient actions, allowing them to maintain and improve their position in the market. Techniques that are based on lexicons, techniques that are based on machine learning, and hybrids that combine these two approaches have all been extensively discussed in the research that has been done on sentiment analysis [1, 2, 3, 4]. The application of a dictionary that has been predefined allows for the determination of the sentiment orientation of a body of textual data using a method that is known as the lexicon-based method. This particular dictionary has a different system for classifying words according to the feelings that they elicit

© 2023 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

in the person reading the dictionary. It achieves a high level of performance when classifying text using its dictionary [5]. SentiStrength 3.0, SentiWrodNet, WordNet, Linguistic Inquiry Word Count(LIWC), Affective Norms for English Words(ANEW), and sentient are lexicon-based tools [6].

A subset of the input data containing the desired output class is provided to the algorithm in order for it to construct the rules. The actual input data to be processed is then provided to the algorithm. This method can reduce the time required to process the data(the test data). There is a dearth of published literature on the use of sentiment analysis in the airline sector, and the majority of this work focuses solely on datasets from the United States. This is despite the fact that there are more and more studies on sentiment analysis being conducted. To the best of the author's knowledge, this is the first paper that uses support vector machines(SVM), a kind of machine learning, to analyse the sentiment of tweets that are relevant to United States Airlines. The researchers were directly encouraged to carry out the current inquiry as a result of the findings of the investigation that came before it. Because of this work, not only will the field of sentiment analysis advance, but it will also help airlines in terms of consumer intelligence and education, which will increase the level of satisfaction experienced by customers. The completion of this task will ultimately lead to an increase in the level of satisfaction experienced by customers. In addition to this, it will place an emphasis on the use of social media analytics not only in the marketing of their products and services but also in the management of risks, the forecasting of enterprises, the examination of competitors, and the development of new products.

2. Introduction of related concepts

The accurate classification of user-generated text into predetermined dichotomies is the primary focus of the expansive field of sentiment analysis, which was named after its primary focus. The primary purpose of sentiment analysis is to achieve this result. Analysis and detection of sentiment can be accomplished with the help of a wide variety of tools and approaches. After being trained with training data, supervised machine learning algorithms are applied. In addition to that, one may make use of a wide variety of additional tools and algorithms. This line of inquiry uses both lexical methods and hybrid tools, which combine lexicon-based classification algorithms and machine learning techniques. Classification can be accomplished with the help of lexical methods by using a corpus that has been annotated with dictionaries. The classification process is carried out by hybrid tools using a corpus that has been annotated with dictionaries. Classification methods that are based on a corpus that has been annotated with the help of a dictionary are known as lexical techniques. In order to carry out the necessary classification tasks, hybrid tools make use of a corpus that has been annotated with dictionaries.

Completing tasks requiring classification is made possible with the assistance of hybrid tools. Support Vector Machine was utilized for several tasks associated with this research project, including sentiment analysis. Because of one of these responsibilities, a Support Vector Machine was necessary(SVM). In the realm of supervised machine learning, support vector machines, more commonly referred to by their acronym SVM, are among the most popularly employed algorithms for determining the polarity of text. Support vector machines can also be referred to using the acronym SVM. The effectiveness of support vector machines(SVM) is evaluated using two datasets of pre-classified tweets, with precision, recall, and F-measure serving as the metrics of choice for conducting a comparative study. Tables and graphs, each in their appropriate format, are utilized here. The SVM was initially mentioned in a formal context by [7] and has since grown to be one of the most well-known applications of supervised machine learning.

It is a typical strategy that has repeatedly surpassed Naive Bayes classifiers and been shown to be extremely successful in numerous aspects of text categorization. In addition, it is highly effective in various text categorization tasks [8]. In addition, it has been demonstrated to be highly effective in various text categorization applications, which is an enormous advantage. It is a method for categorizing texts that has recently received much attention due to its effectiveness in various contexts. This is because it can be utilized to categorize texts. Together with both sets of data, precision and recall statistics, as well as the F-measure, are used to evaluate the support vector machine's(SVM) performance. The practice of analyzing the feelings conveyed in the text is currently trendy. The process of extracting data from social networking sites and analyzing it is currently being automated by researchers. In the research paper [9], the authors proposed a method for obtaining pre-labelled data from Twitter to train SVM classifiers.The proposed method was found to be accurate 85 percent of the time based on the classifier evaluation.

In another piece of research, J48 and MLP were compared across five different datasets. The precision, recall, F-measure, and ROC Area were the metrics used to ascertain the accuracy level [10]. MLP outperformed the other companies in the industry. In addition, the findings showed that Neural Networks could be utilized for classification purposes. A novel method [11] was utilized to classify tweets as either positive or negative. They talked about conducting sentiment analysis on Twitter through remote machine learning. Tweets with annotations served as the basis for the training data. According to the authors, Naive Bayes, Maximum Entropy, and SVM can achieve an accuracy of more than 80% when applied to tweets containing emoticons. Accuracy in classification preprocessing was the primary focus of the study's investigation. In another piece of research, the content of Arabic Twitter was dissected. They used NB and SVM to analyze one thousand tweets to determine polarity [12]. The research found that using feature vectors improved the performance of machine-learned classifiers. In the training data, we found opinions that contradicted one another, repeated opinions, and opinion spamming. This may result in a loss of precision; the study's authors classified Arabic tweets using a combination of Nave Bayes, Decision Trees, and Support Vector Machine. In this study, Arabic tweets were categorized by employing TF-IDF and stemming. They assessed the accuracy of three algorithms by using a single dataset's precision, recall, and f-measure.

3. Methodology

3.1. Materials

This research makes use of datasets of tweets that have been previously labelled [13]. This dataset includes tweets that are related to US-Airline. There are 2363 positive tweets, 3099 neutral tweets, and 9178 negative tweets for six US airlines, for a total of 14640 tweets.

Table 1. Tweets dataset for 6 US airlines.

CLASS	TWEETS
NEGATIVE	9178
NEUTRAL	3099
POSITIVE	2363
TOTAL	14640

3.2. Preprocessing

The preliminary data processing before it is used in the classification procedure is critical. At this point, the dataset will be normalized and made ready for the classification algorithm. This ensures that the algorithm runs without hiccups and produces valuable results in the shortest possible time [8]. According to several studies, the preprocessing parameters include the TF-IDF, the Stemmer, the stopwords Handler, and the tokenizer, among other things [1, 14, 15].

3.3. Classification

In this phase, the normalized data is subjected to the support vector machine(SVM) for classification purposes, and the outcomes are reported. Any supervised machine learning algorithm can be subjected to performance analysis by using pre-classified data as test data and comparing the output polarities to the polarities used during the classification process. A data set composed of pre-marked tweets is used as input data. The precision, recall, and f measure metrics are used to assess the results' correctness.

3.4. Results

This section analyzes both datasets' SVM results and compares them using a variety of metrics. This study uses three separate evaluation parameters—precision, recall, and F measure—to enable comparison.

The TP rate and FP rate can be used to calculate the precision in the following ways:

$$\mathrm{TP}$$

$$Precision = \frac{}{(TP + FP)}$$

The notation TP is applied to sentences that have been correctly classified, whereas the notation FP is used for sentences that have been misclassified.

The formulas for calculating recall are as follows:

$$Recall = \frac{\mathrm{TP}}{(TP + FN)}$$

FN indicates a sentence that is not classified, whereas TP indicates a sentence that is correctly classified (as explained above).

The formulas below is used to determine the F-measure :

$$F - measure = \frac{\mathrm{Precision}\times\mathrm{Recall}\times 2}{(Precision + Recall)}$$

According to the findings, the accuracy, precision, and recall, as well as the F1, come in at respective percentages of 91.8 percent, 91.3 percent, 82.8 percent, and 86.9 percent.

Table 2. Accuracy, precision, recall and F1 score.

ACCURACY	91.8%
PRECISON	91.3%
RECALL	82.8%
F1	86.9%

4. Conclusion

The aim was to know how well the Support Vector Machine(SVM) did when it was given the task of analyzing people’s feelings, so we conducted this research. In order to evaluate how well SVM works, we used a single pre-classified dataset that was derived from tweets. This dataset was our test subject. This data set included tweets that were relevant to six different airlines with headquarters in the United States. Python is the programming language that is utilized for the purposes of performance analysis and benchmarking. The precision, recall, and f-measure metrics are utilized in order to perform an accuracy analysis on the results. According to the investigation, the dataset had an accuracy of 91.8 percent, a precision of 91.3 percent, and a recall of 82.8 percent. In addition, the value of f1 can be expressed as 86.9 percent at the present time. In the following section, the total set of results is presented. The findings make it abundantly clear that the performance of SVM is contingent on the dataset that is provided as input. Further research is required to understand how SVM and other machine learning approaches are affected by performance, and more research needs to be done. For the purpose of this study, a wide variety of different large datasets should be utilized. The findings of this research project have the potential to serve as a benchmark for subsequent comparative research after it has been completed.

Acknowledgment

I would like to express my most heartfelt gratitude to each and every instructor and professor at my university. They have been of incredible assistance to me throughout the entire process of developing my thesis, and I simply cannot thank them enough. In addition, I would like to express my gratitude to both of my parents as well as to each and every one of my friends for their constant support and encouragement. Without their extraordinary intelligence and politeness, I never would have been able to finish my thesis, which is a major accomplishment for me.

References

[1] Ahmad, M., & Aftab, S. (2017). Analyzing the Performance of SVM for Polarity Detection with Different Datasets. International Journal of Modern Education and Computer Science(IJMECS), 9(10), 29- 36.

[2] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1–2), 1-135.

[3] Saif, H., He, Y., Fernandez, M., & Alani, H. (2016). Contextual semantics for sentiment analysis of Twitter. Information Processing & Management, 52(1), 5-19.

[4] Ahmad, M., Aftab, S., Ali, I., & Hameed, N. (2017). Hybrid Tools and Techniques for Sentiment Analysis: A Review. Int. J. Multidiscip. Sci. Eng, 8(3).

[5] Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

[6] Ahmad, M., Aftab, S., Muhammad, S. S., & Waheed, U. (2017). Tools and Techniques for Lexicon Driven Sentiment Analysis: A Review. Int. J. Multidiscip. Sci. Eng, 8(1), 17-23.

[7] Cortes, C., & Vapnik, V. (1995). Support vector machine. Machine learning, 20(3), 273-297

[8] Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics.

[9] Zgheib, W. A., & Barbar, A. M. (2017). A study using support vector machines to classify the sentiments of tweets. International Journal of Computer Applications, 975, 8887.

[10] Arora, R. (2012). Comparative analysis of classification algorithms on different datasets using WEKA. International Journal of Computer Applications, 54(13).

[11] Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N project report, Stanford, 1(12), 2009.

[12] Shoukry, A., & Rafea, A. (2012, May). Sentence-level Arabic sentiment analysis. In 2012 international conference on collaboration technologies and systems (CTS) (pp. 546-550). IEEE.

[13] Twitter US Airline Sentiment from https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment.

[14] Altawaier, M. M., & Tiun, S. (2016). Comparison of Machine Learning Approaches on Arabic Twitter Sentiment Analysis. International Journal on Advanced Science, Engineering and Information Technology, 6(6), 1067-1073.

[15] Isa, D., Lee, L. H., Kallimani, V. P., & Rajkumar, R. (2008). Text document preprocessing with the Bayes formula for classification using the support vector machine. IEEE Transactions on Knowledge and Data engineering, 20(9), 1264-1272.