Sentence Level Sentiment Identification and Calculation from News Articles Using Machine Learning Techniques

Vishal S. Shirsat, Rajkumar S. Jagdale and Sachin N. Deshmukh

Abstract Sentiment analysis is a widely used phenomenon for analyzing online user responses to infer collective response and it is used in various applications. Negation is a very common morphological creation that affects polarity. This research paper focuses on sentence level negation identification from news articles this work uses online news articles Data from BBC news. Results are analyzed using Machine Learning Algorithms like Support vector Machine and Naïve Bayes. Support Vector Machine achieves 96.46% accuracy and Naive Bayes achieves 94.16%.

Keywords Sentiment analysis · Support vector machine · Naïve Bayes · Machine learning algorithm · Negation identification

1 Introduction

Sentiment Analysis is an application of Natural Language Processing, computational linguistics and text analytics which identifies and extract subjective information in source materials such as product reviews, chats and discussions [1]. Sentiment analysis determines the inclination of a correspondent through the contextual polarity of their language or writing, their attitude which may be pretended in their own judgment, emotional state of the substance, and otherwise the state of any emotional communication they are using to affect a reader. It is demanding to define a person's state of mind on the topic they are collaborating about. This information can be mined from several data sources from texts, tweets, blogs, social media, and news

V. S. Shirsat (✉) · R. S. Jagdale · S. N. Deshmukh
Department of Computer Science and IT, B. A. Marathwada University, Aurangabad, India
e-mail: vss.csit@gmail.com

R. S. Jagdale
e-mail: rajkumarjagdale@gmail.com

S. N. Deshmukh
e-mail: sndeshmukh@hotmail.com

B. Iyer et al. (eds.), *Computing, Communication and Signal Processing*,
Advances in Intelligent Systems and Computing 810,
https://doi.org/10.1007/978-981-13-1513-8_39

articles [2]. News articles and web blogs are one of the most essential platforms that permit users to express their personal opinion about the several topics. Basically Sentiment analysis covers a big part of, computational linguistics, natural language processing, and text mining. Generally, the aim of sentiment analysis is to finding the polarity of opinion. In statistical way, sentiment analysis methods are based on frequency of positive and negative words. Many researchers have identified the ways of accounting for several other features of content, for example structural aspects.

2 Related Work

Nowadays online news and web blogs have become an important source of information from news websites. People share their thoughts, feelings in the form of news articles and web blogs. This increase in the amount of online opinion-related textual information has led to the rapid development of the field of sentiment analysis. Pang et al. [3] have used machine learning approaches to determine the accuracy of classification from documents. Experiments were performed on movie data and it was concluded that the machine learning techniques are always better than human made baseline for sentiment analysis. The work involves intensifiers analysis to extract exact sentiments. Machine Learning approaches like Naïve Bayes, Maximum entropy and support vector machines classification techniques are used. As an inference, it is concluded that machine learning techniques are better than human baselines for sentiment classification. Mohammad et al. [4] defined a method to increase the scope of sentiment lexicon and includes the Identification of separate words and multi-word expressions. Lexicon and a list of affixes are used here. This method can be implemented using antonym generation or lexicon based. For antonym generation Hand-crafted rules were used. Lexicon approach is based on the word list which defines if a paragraph has more negative words than the positive ones and accordingly polarity of the paragraph is decided. Turney [5] has used semantic mining for binary classification and part of speech (POS) tagging. He worked on document level and review level sentiment analysis. Shoukry [6] shows an application for Arabic language sentiment analysis and performed a sentiment classification for Arabic tweets. The collected tweets are examined to provide their polarity. Their study proposed hybrid system that used all the identified features from the ML approach, and the sentiment lexicon from the SO approach, resulting in an accuracy and recall of 80.9%, while its precision and F-measure is 80.6%. Alexandra Balahur [7] stated the importance of the tasks in three levels. The work separated good and bad news content. Ding and Melville [8, 9] focused on machine learning approaches to train classifier. Lexicon dictionary based method depends on corpus or list of words having certain polarity. An algorithm pointed out the dictionary words and calculated the weight accordingly.

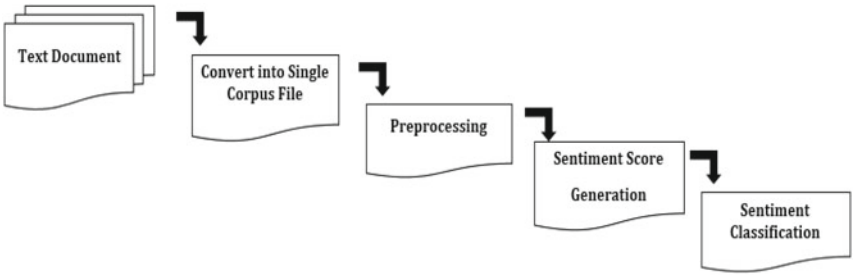

Fig. 1 Proposed methodology

3 Preprocessing

Sentiment analysis is the process which identifies or expresses the polarity of the text data. Basically, Sentiment analysis has been categorized in three levels: first Document level sentiment analysis, here whole document is to be to find the polarity of that document. For example, if one text file contains reviews of only one product, then the system calculates polarity of complete text in the document. Thus the document expresses opinion on a single entity and is not applicable on multiple product reviews [10]. In sentence level, every sentence is processed and analyzed to determine the polarity. Aspect level sentiment analysis helps to realize find out sentiment on objects and their features [11].

The preprocessing of the dataset and preparing the text for classification is an important task. The work presented in the paper uses BBC News Article Dataset. Online text contains irrelevant text such as HTML tags, scripts, and advertisements. Preprocessing plays a very vital role in text mining methods and applications. It is a first and foremost steps which will help to cleaning a data and increasing the data sparsity and substantially shrinking the feature space [12]. There is no impact on the general orientation in word level sentiment analysis [13]. It also helps for to cleaning and preparing the text for classification. Basically online and offline data is having huge amount unwanted information which does not contain any wanted information from the data. Removing the unwanted information and meaningless information, from the text data such type of things comes under in preprocessing stages (Fig. 1).

4 Machine Learning Algorithms

Naïve Bayes classifier uses Bayes Theorem, which finds the probability of an occurrence given the probability of another occurrence that has already occurred. NB classifier does particularly well for problems which are linearly separable and even for problems which are nonlinearly separable it perform reasonably well [14].

$$P(A|B)\frac{P(B|A)P(A)}{P(B)} \tag{1}$$

Support vector machine is non probabilistic algorithm which is used to separate data sequentially and Non-sequentially [15]. It is basically used for text classification and get a good performance in high-dimensional feature space. Support Vector Machine algorithm denotes of the instances points in space, mapped so that the instances of the different classes are separated by a clear margin as extensive as possible [16].

5 Experimental Results

Our proposed methodology uses mainly five steps. In first steps performs data cleaning and removes URL, Stop words, Punctuation, Strip white space and Numbers from the data. The step of number removal is important a number hardly represents the sentiments and hence not useful. Then the next step performed is to convert the whole document into lower case to have uniformity and then the stemming. Stemming is used for to change word a root form of the word for example. "Education", "Educated", "Educating" will be converted into single word, i.e., educate. After preprocessing the dataset, we need to determine the Term Document Matrix which describes the frequency of terms that occur in the processed dataset. Rows in the dataset are considered as a collection and column considers as a related terms. This is achieved by using "dtm" function in TM package of R. After preprocessing, there is a need of sentiment score generation with the help of positive and negative dictionary. Each word in the dataset will be compared with the dictionary word to determine whether it is positive or negative. Further, Naïve Bayes and Support Vector Machine algorithms are used for the classification purpose and accuracy is estimated.

Table 1 below shows the category of the article and the count of neutral, positive and negative word in it. This work uses Bing Liu dictionary which contains 2006 positive word and 4783 negative word. The results are as following (Fig. 2).

For the classification two machine learning algorithms Naïve Bayes and Support Vector Machine are used, and Accuracy, Precision and F-Score of the Data is

Table 1 Category wise document polarity

Sr No	Name of category	Total	Positive	Negative	Neutral
1	Business	510	262	214	34
2	Entertainment	401	136	244	21
3	Politics	417	210	190	17
4	Sport	511	151	327	33
5	Tech	401	136	244	21

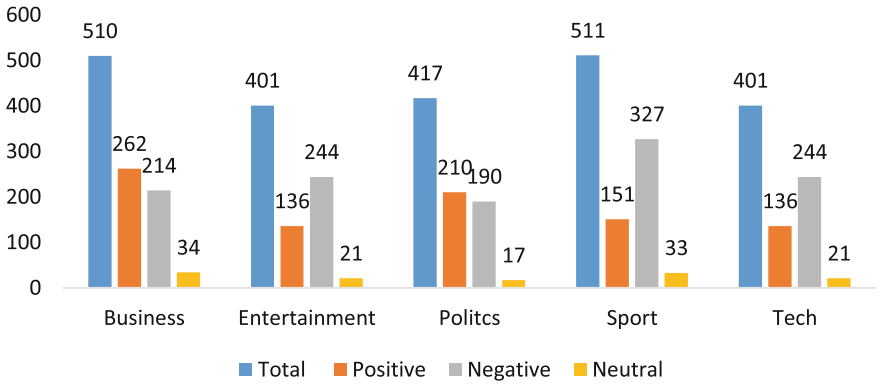

Fig. 2 Graphical representation of document polarity

Table 2 Comparative analysis of Naïve Bayes and SVM

Dataset	Naïve Bayes			SVM		
	Accuracy	Precision	F-score	Accuracy	Precision	F-score
Business	92.63	89.76	91.32	82.60	79.67	89.34
Entertainment	96.46	94.80	97.33	69.91	68.22	84.39
Politics	93.33	88.88	93.33	94.16	89.06	94.21
Sport	93.00	90.74	95.14	69.23	69.01	81.66
Tech	96.46	94.80	97.33	69.91	68.22	81.11

estimated. As per experimentation, Naïve Bayes achieves 96.46% accuracy for Entertainment category and lowest accuracy for Business category, i.e., 92.63%. Similarly, with the Support Vector Machine 94.16% accuracy is achieved for Politics Category and Lowest Accuracy for Sport Category, i.e., 69.01%.

6 Conclusions and Future Scope

The work has been accomplished to find the polarity of the news articles. The result shows the category wise document polarity. The work is based on the Dictionary based approach with machine learning techniques. From above experimentation, it can be said that Naïve Bayes gives better results than Support Vector Machine as shown in Table 2. Future work will focus on other Classification techniques on other data related to online news articles and web blogs.

References

1. Roebuck, K.: Sentiment Analysis: High-Impact Strategies What You Need to Now: Definitions, Adoptions, Impact, Benefits, Maturity. Vendors, Emereo Publishing, 05 Nov 2012
2. Pooja, P., Sharvari, G.: A survey of sentiment classification techniques used for indian regional languages. Int. J. Comput. Sci. Appl. **5**(2) April 2015
3. Bo, P., Lillian, L., Shivakumar, V.: Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79–86 (2002)
4. Mohammad, S., Dorr, B., Dunne, C.: Generating high-coverage semantic orientation Lexicons from overly marked words and a thesaurus. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp. 599–608 (2009)
5. Turney, P.: Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the Association for Computational Linguistics, pp. 417–424, Philadelphia (2002)
6. Shoukry, A.: Collaboration Technologies and Systems (CTS). In: International Conference technologies and Systems, 21–25 May, pp. 546–550 (2012)
7. Alexandra, B., Ralf, S.: Rethinking Sentiment Analysis in the News, Theory to Practice and back‖, European Commission, Joint Research Centre, Department of Software and Computing Systems, University of Alicante, WOMSA, pp. 1–12 (2009)
8. Ding, X., Liu, B., Yu, P.: A holistic lexicon-based approach to opinion mining. In: Proceedings of the International Conference on Web Search and Web Data Mining, pp. 231–240. ACM (2008)
9. Melville, P., Gryc, W., Lawrence, R.: Sentiment analysis of blogs by combining lexical knowledge with text classification. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp. 1275–1284 (2009)
10. Emma, H., Xiaohui L., Yong S.: The role of text pre-processing in sentiment analysis. Procedia Comput. Sci. Elsevier, **17**, 26–32 (2013) [14] Tetlock, P., Saar-Tsechansky, M., Macskassy, S.: More than words: quantifying language to measure firms fundamentals. J. Financ. **63**(3), 1437–1467 (2008)
11. Bing, L.: Sentiment Analysis and Opinion Mining, Apr 22 (2012)
12. Melville, P., Gryc, W., Lawrence, R.: Sentiment analysis of blogs by combining lexical knowledge with text classification. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1275–1284. ACM (2009)
13. Jagdale, R.S., Shirsat, V.S., Deshmukh, S.N.: Sentiment analysis of events from twitter using open source tool. Int. J. Comput. Sci. Mob. Comput. **5**(4), pp. 475–485 (2016)
14. Ye, Q., Zhang, Z., Law, R.: Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. Expert Syst. Appl. **36**, 6527–6535 (2009)
15. Bhumika, M., Jadav, V., Vaghela, B.: Sentiment analysis using support vector machine based on feature selection and semantic analysis. Int. J. Comput. Appl. **146**(13) (2016)
16. BholaneSavita, D., Deipali, G.: Sentiment analysis on twitter data using support vector machine. Int. J. Comput. Sci. Trends Technol. **4**(3) (2016)