Intelligent Systems and Computer Technology
D.J. Hemanth et al. (Eds.)

doi:10.3233/APC200124

Polarity Detection on Real-Time News Data Using Opinion Mining

Boppuru Rudra Prathap [a,1], Sujatha A K [b], Chandragiri Bala Satish Yadav [c], Mummadi Mounika [d]
[a,b,c,d] *CHRIST (Deemed to be University), Bangalore, India*
[a] *Email:boppuru.prathap@christuniversity.in*
[b] *Email:sujatha.ak@christuniversity.in*
[c] *Email:chandragiri.yadav@btech.christuniversity.in*
[d] *Email:mounikamummadi98@gmail.com*

Abstract: Sentimental Analysis or Opinion Mining plays a vital role in the experimentation field that determines the user's opinions, emotions and sentiments concealing a text. News on the Internet is becoming vast, and it is drawing attention and has reached the point of adequately affecting political and social realities. The popular way of checking online content, i.e. manual knowledge-based on the facts, is practically impossible because of the enormous amount of data that has now generated online. The issue can address by using Machine Learning Algorithms and Artificial Intelligence. One of the Machine Learning techniques used in this is Naive Bayes classifier. In this paper, the polarity of the news article determined whether the given news article is a positive, negative or neutral Naive Bayes Classifier, which works well with NLP (Natural Language problems) used for many purposes. It is a family of probabilistic algorithms that used to identify a word from a given text. In this, we calculate the probability of each word in a given text. Using Bayes theorem, they are getting the probabilities based on the given conditions. Topic Modeling is analytical modelling for finding the abstract of topics from a cluster of documents. Latent Dirichlet Allocation (LDA) is a topic model is used to classify the text in a given document to a specified topic. The news article is classified as positive or negative or neutral using Naive Bayes classifier by calculating the probabilities of each word from a given news article. By using topic modelling (LDA), topics of articles are detected and record data separately. The calculation of the overall sentiment of a chosen topic from different newspapers from previously recorded data done.

Keywords. Polarity detection, LDA, Sentiment analysis, Fake news detection.

1. Introduction

Nowadays, the amount of data generated is large and reading and understanding that data is an endless task. However, continuous attempts and analysis in this field have led to the process of automated to some extent. The further plan is to atomize this PROCESS. This is true in the case of online platforms and social network sites (SNSs) where people can share the information which can go viral in a few hours. According to an approach to the detection of news, whether it is positive or negative, can be done in two ways. The first approach is to use the human ability to describe the truth from given information, which is using a human understanding to determine whether the news is positive or negative. The second approach is to use the algorithm approach to

[1] Boppuru Rudra Prathap, CHRIST (Deemed to be University), Bangalore, Karnataka, India.
E-Mail: boppuru.prathap@christuniversity.in.

identify whether the news is positive or negative in this article. The introduction to Sentimental analysis explained. Sentimental analysis mostly used application to process the data and extract the information. It is the process of collecting and analyzing the data based on the user's feelings. It performed using various machine learning techniques and NLP strategies (Natural Language Processing). It can find whether the given text positive, negative or neutral. Web Scraping is a technique used to extract a colossal volume of data from different websites. The data derived from any online website does not have a facility to save a copy. So to get the data, it is needed to copy and paste the data manually. So to overcome this web scraping technique used where this task is automated. This derivation of data used to get information from available resources, which will renew the datasets in real-time. These resources can be online news websites. Polarity detection can done using Naïve Bayes Classifier. It gathers the probabilities for each class and predicts that the likelihood of a given data belongs to a particular class. Naive Bayes has been studied extensively since the 1950s. It is a method that had introduced for text categorization, which is used to check that the documents belong to one category or another in which the word frequency is used as a feature. With preprocessing, it is possible in this domain with more advanced methods, including support vector machines (SVM), resulting in improved accuracy. Since topic modelling can determine concealed semantic structures, researchers used sentiment analysis models based on topic models. These models had successfully applied to long texts, but analysis for short text is a challenging task because of the sparsity of features in short paragraphs. Polarity Detection is a method in which the extracted new articles from different news websites is determined, whether positive or negative, using sentimental analysis. In this, the preprocessing of data can be done by removing special characters and changing unstructured data to structured data. Naïve Bayes classifier used for calculating the probabilities of each news article and based on these probabilities it determines whether the given news article is positive, negative or neutral. LDA (Latent Dirichlet Allocation) is used for topic modelling to find the topic of the news articles and record the data separately. After finding all the topics from news articles, the sentiment of each topic is calculated from different newspapers and then determining whether which newspapers supporting or not supporting on a particular topic. In this research, the polarity of the news article is calculated and then compared the calculated polarity with the polarity from the news list API to check the accuracy.

2. Literature Survey

2.1. News Analysis

B. Wen, S. Duan, B. Rao and W. Dai [8] have discussed word sentimental orientation identification, which falls into two classes: The shortcomings of this paper are their research based on individual words which ignore the context on words sentimental orientation. A. Kottwani et al. [11] claiming that Text summarization combines the process of term frequency, topical analysis and POS tagging. P. Shah and N. P. Desai [12] discuss that nowadays, textual is increasing day by day and is available in many different languages. It is becoming difficult to read all the content. So to overcome this Text summarization used. In this paper, they are using a technique called Automatic Text Summarization, which summarizes the large text into small text by

extracting relevant content. In this paper, they are discussing automatic text summarization for various Foreign and Indian Languages. Different sequences of features work individually for separate content. It is challenging to create a single summarization technique for peculiar content. F. B. Ashraf et al. [14] discuss that as online news articles have with the advancement of information and technology, some of the news is brutal, so to identify and categorize these news articles automatically, they used sentimental analysis and opinion mining. The author used sentimental analysis to detect whether the given news article is positive or negative by using the sentence identification phase. This process is also helpful in determining the review of the products for large scale companies. T. Yamada [15] proposes a topic detection method using a topic model for a Japanese newspaper and to visualize the time change of the detected topics. In this, they used Mainichi newspapers from 2010 to 2015. There are almost six hundred news articles. They extracted nouns as the keywords of the given text. They used LDA to detect the topic of the given text, where LDA is one of the topic models. In this paper, they are discussing the earthquake topics. In order to grasp the topic, they visualized the change of the frequency of the occurrence of the topic and top words on a monthly basis. Analyzing topics by region helps to grasp the situation fluctuation in the region. In this paper, they discussed the topic detection using LDA and the visualization to analyze the change in time series. LDA is one of the unsupervised learning of the model and LDA classifies data using a classifier. The supervised learning uses the classification indicator to classify the data.

2.2. Fake News Analysis

In the modern years, the growing of falsity has been increasing on the internet and has been influencing the political and social existence. Fake news can be analyzed by using different machine learning algorithms and Artificial Intelligence. Fake news can be identified in online news, news blogs, Facebook and Twitter where a user posts fake news. E. Tacchini et al. [3] discussed that the propagation of fake news needs the automation of fake news detection. Machine Learning methods can be used for this purpose. In this paper, they used fake news detection by combining social content and news content features and secondly, it is implemented in chatbot and Facebook messenger and validating it with a real-world application. S. Padmaja et al. [10] have explained that opinion mining can be done on subjective text types such as text like product or movie reviews. In this paper, it describes that opinion mining is crucial than topic sentiment classification. M. Granik and V. Mesyura [1] and A. Jain and A. Kasbe [2] discussed that fake news detection methods done using one of the artificial intelligence algorithms – Na¨ıve Bayes classifier which implemented against a dataset of Facebook posts. The result of these papers shows that even many more artificial intelligence techniques can be used to handle the problem. In [2], they implemented a new concept called web scraping, which can be used to update the data sets so that the results can be made more precise and accurate. E. Tacchini et al. [3] used fake news detection methods by combining both social content and news content features, and it implemented in chatbot and Facebook messenger and validating it with a real-world application. S. Helmstetter and H. Paulheim [4] discussed weekly supervised way, which accordingly detects large training datasets, including hundreds of thousands of tweets. This author builds a large training dataset to predict the integrity of a tweet rather than the certainty of the tweet itself. The major objection is acquiring extensive training data since manual verification of tweets as not fake is a valuable task. C.

Buntain and J. Golbeck [5] developed a method for automatizing fake news detection on twitter by predicting accuracy in two reliability twitter datasets: PHEME, a dataset of possible rumours in twitter, CREDBANK, a group collection dataset of certainty determination for events in Twitter. This method used to Twitter content collection from fake news datasets. Results show that the accuracy prediction model classifies the twitter fake news stories. Additionally, accuracy models achieved from group collection works on experienced journalists in differentiating possible fake twitter threads and feature analysis. S. Wankhede et al. [6] used twitter, which is one of the most used social media, which has more than 200 million tweets per day. In this paper, they used the Hidden Markov Model and N-gram method for Correction of Tweets and Spell Checking and also Emoji Sentiment Ranking method, which used to classify sentiment aligning of emojis such as positive, negative and neutral. M. K. Bedi et al. [7] are doing sentimental analysis on twitter data. In this paper, they used the Fuzzy classifier and na¨ıve Bayes to evaluate tweets into positive, negative and neutral. M. Rathi, A. Malik, D. Varshney [9] used Twitter, which is a microblogging website where users share data in the form of tweets. In this, they are using all machine learning techniques to increase efficiency and to improve the classification results of the sentimental analysis. In this, they merged SVM (Support Vector Machine) with a decision tree to provide better-evaluating results in accuracy and f-measure. S. Krishnan and M. Chen [13] discuss how social media can be used for the spread of rumours and fake news for financial and political benefits. Considering the effect of false news, they decided to detect false information to prevent it from spreading. This author used statistical analysis of twitter user account, reverse image searching, cross verification for fake news and data mining.

This area introduces an approach to find the sentiment of a given text and determine whether the given text is positive, negative or neutral. The first literature survey conducted on newspaper analysis in which it led us to find out that news analysis can be conducted to determine Crime news, Fake news and Text summarization. Nowadays as data online is becoming vast and any kind of shared information is going viral within no time polarity detection method is used to automate the process of analyzing the data by using sentimental analysis and using Na¨ıve Bayes algorithm to determine the probabilities of each word termed as positive, negative and neutral. According to the Literature LDA is used for topic modelling to detect the topics of the news articles from different newspapers.

3. Methodology

The main objective of the work is to perform sentimental analysis on different newspapers. To achieve this objective Machine Learning algorithm is used on the newspapers by using predefined positive and negative words datasets. The following steps had followed to achieve the main objective. 1.A thorough study of existing techniques and libraries for performing data analysis in python.2.Collection of datasets from online.3.Collection of news articles from different newspaper websites.4.Pre-processing of the data collected is performed so that it can be used for analysis.5.Calculating the probabilities of the articles termed as positive, negative or neutral.6.Using topic modeling finding the occurrences of each topic from different newspapers.7.Storing the data in the relevant data frame.8.Calculating the sentiment of each topic to find which newspaper is supporting the topic.

3.1. Proposed Architecture:

In Figure-1 First we input the URLs from online newspaper websites and extract the data by using web scrapping technique (To extract large amounts of data).Then we pre-process the extracted data by removing special characters, double spaces, stop removal words.The pre-processed data is used to do sentimental analysis using na¨ıve Bayes classifier and determining the probabilities of the each word in the given news article expressed in terms of positive, negative or neutral.After getting sentiments of the given new articles we compare them with each other for the supportiveness. In the Figure-2 example articles 'A' and 'B' have the same sentiment and article 'C' has the different so it means that 'A' and 'B' are supporting the same news and 'C' is against it. The data is stored and used in topic modeling to find the topic of the news article by calculating the sentiment of each word.

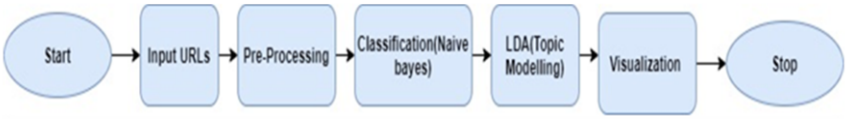

Figure-1 Proposed Architecture.

In the whole process of the project, python is running on PyCharm IDE to connect with server. Pre-Defined Datasets are extracted from online and the datasets are stored separately. Articles are also extracted from different newspaper websites using requests library. The URLs are gathered from different newspapers and Inputting the URLs to the console using requests Library in Python. The response data is taken from requests library and extracting the data from URLs using Web scrapping technique. Beautiful Soup is used to extract the data. Later the title and data are stored separately.

	Article A	Article B	Article C
Positive	0.8946	0.7532	0.3156
Negative	0.3456	0.4638	0.7894
Overall	Positive	Positive	Negative

figure-2 Example of article wise probabilities

4. Results and Discussion

The sentiment analysis is efficient on categorizing the online news articles based on the polarity detection using Naive Bayes Classifier and later the topic is detected using topic modeling. In this predefined datasets are used to calculate the probabilities of the words in the given news articles. Firstly using Requests library the URLs are extracted from online websites and the data is scrapped from those online news articles using Beautiful Soup and the expected output is the data from the online news articles should be extracted and displayed in console and stored in CSV file. In this project news articles are compared related to same topic from three different newspapers. Stored data is preprocessed by removing stop words, special characters, URLs and the

data is processed data is stored separately by separating the words using comma. The expected output is the data should not have any extra characters apart from the keywords extracted and a list is formed using these keywords. Using this keywords list and datasets extracted from online the probability of the each keyword is determined by using Naive Bayes classifier. Based on the calculated probabilities if the news article have more good words then it will be assumed as positive ,more bad words it will be assumed as negative and more neutral it is assumed to be neutral. then calculated probability is compared with the probability from the API. After this a graph is plotted using matplot library to compare the values of the calculated probability and API probability. Then graphs are plotted to determine which newspaper has more positive, negative and neutral sentiments. Later topic modeling is used to detect the topic from the processed news articles so that the overall probability can be calculated and then determine which topic has more frequency and find the topic of the newspapers. In this LDA is used to detect the topic. In this the frequency of each keyword in a given document is checked and the topic is determined by analyzing the sentiment of all keywords from the given documents.

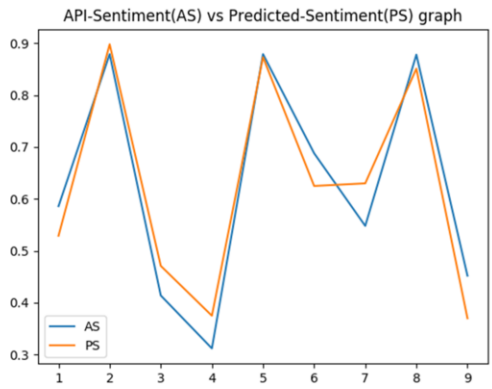

Figure- 3 API Vs Predicted Polarity Comparison

In Figure-3 the comparison of calculated polarity and the API polarity of the news articles is shown where AS represents API polarity and PS represents calculated polarity.

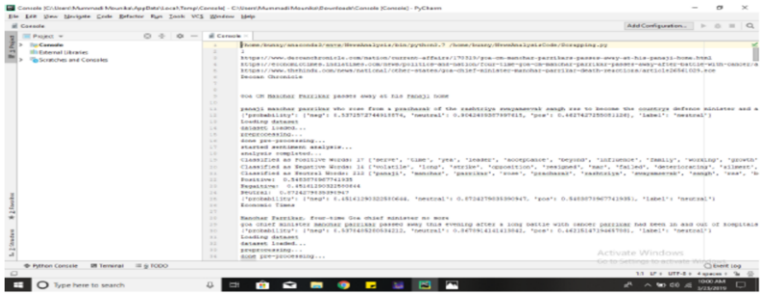

Figure-4 Sentimental Analysis Of News Articles

In Figure-4 result of the overall sentimental analysis of the news data is shown. First the probabilities are extracted from API. The datasets are loaded. Later data is preprocessed and sentimental analysis is carried out. In this project three different newspapers (Hindu, Deccan Chronicle, and Economic Times) on same topic are taken.

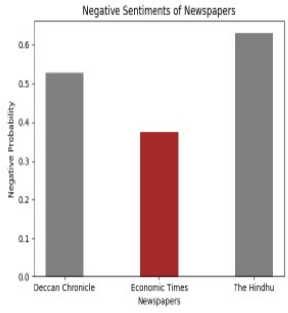

FIGURE-5
Negative Polarity for Newspapers

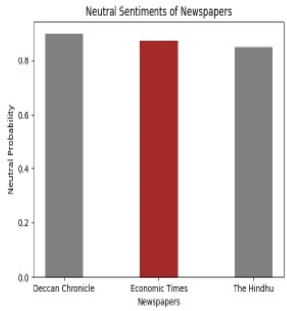

FIGURE-6
Neutral Polarity for Newspapers

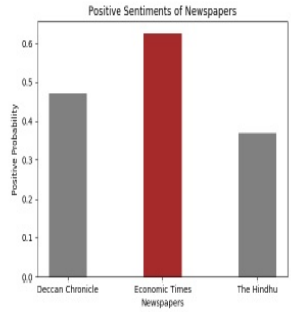

FIGURE-7
Positive Polarity for Newspapers

Figure-5 it is comparing the negative polarity of all the three newspapers and finally it is showing that The Hindu newspaper has more negative polarity of all the newspapers.Figure-6 it is comparing the neutral polarity of all the three newspapers and finally it is showing that Deccan Chronicle has more neutral polarity of all the newspapers. Figure-7 it is comparing the positive polarity of all the three newspapers and finally it is showing that Economic Times has more positive polarity of all the newspapers. In proposed system three news articles on similar topic are compared from three different newspapers to determine the perspective of the newspapers on the given topic using Naïve Bayes Algorithm. In this project Naïve Bayes Classifier is used instead of SVM (Support Vector Machine) because Naïve Bayes is a probabilistic model and is independent of features of a given problem whereas SVM interacts with the features of the other models. In this AI technique which is LDA is also used to detect the topic of the given analyzed data. LDA is used because it is unsupervised learning technique which does not need any classification indicator to classify the data. Sometimes, LDA also have classification indicator to classify the data.

5. Conclusion

In this paper, Naive Bayes used, which can be applied to a huge amount of data to determine whether the given data is positive, negative and neutral — established how to gather the required data for sentiment classification and the cleaning that is necessary with such data. Naive Bayes Classifier is used on collected data and conducted sentiment analysis and have found this process to be successful. LDA is used in this to detect the topic by applying sentimental analysis on the keywords from previously analyzed data from all the documents. As LDA is an unsupervised learning technique, it does not need any classification indicator to classify the data. In the proposed system, three newspapers are taken from different websites to extract the

data. In the future, this process can increase to multiple papers. Our system runs efficiently and accurately with the increase in data. This project can further extend to extracting data from multimedia and news channels. So Topic modelling can be more relevant with the increase in sources. Multiple works can incorporate into this project. One among them is to establish a Fake News detection system based on previous knowledge, which helps in detecting whether the news is fake or real.

References

[1] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kiev, 2017, pp. 900-903.
[2] A. Jain and A. Kasbe, "Fake News Detection," 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), Bhopal, 2018, pp. 1-5.
[3] M. L. Della Vedova, E. Tacchini, S. Moret, G. Ballarin, M. DiPierro and L. de Alfaro, "Automatic Online Fake News Detection Combining Content and Social Signals," 2018 22nd Conference of Open Innovations Association (FRUCT), Jy- vaskyla, 2018, pp. 272-279.
[4] S. Helmstetter and H. Paulheim, "Weakly Supervised Learning for Fake News De- tection on Twitter," 2018 IEEE/ACM International Conference on Advances in So- cial Networks Analysis and Mining (ASONAM), Barcelona, 2018, pp. 274-277.
[5] C. Buntain and J. Golbeck, "Automatically Identifying Fake News in Popular Twit- ter Threads," 2017 IEEE International Conference on Smart Cloud (SmartCloud), New York, NY, 2017, pp. 208-215.
[6] S. Wankhede, R. Patil, S. Sonawane and P. A. Save, "Data Preprocessing for Effi- cient Sentimental Analysis," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, 2018, pp. 723-726.
[7] R. Mehra, M. K. Bedi, G. Singh, R. Arora, T. Bala and S. Saxena, "Sentimental analysis using fuzzy and naive bayes," 2017 International Conference on Comput- ing Methodologies and Communication (ICCMC), Erode, 2017, pp. 945-950.
[8] B. Wen, S. Duan, B. Rao and W. Dai, "Research on Word Sentimental Classification Based on Transductive Learning," 2015 8th International Symposium on Computa- tional Intelligence and Design (ISCID), Hangzhou, 2015, pp. 153-156.
[9] M. Rathi, A. Malik, D. Varshney, R. Sharma and S. Mendiratta, "Sentiment Anal- ysis of Tweets Using Machine Learning Approach," 2018 Eleventh International Conference on Contemporary Computing (IC3), Noida, 2018, pp. 1-3.
[10] S. Padmaja, S. S. Fatima and S. Bandu, "Analysis of sentiment on newspaper quo- tations: A preliminary experiment," 2013 Fourth International Conference on Com- puting, Communications and Networking Technologies (ICCCNT), Tiruchengode, 2013, pp. 1-5.
[11] H. Gupta, A. Kottwani, S. Gogia and S. Chaudhari, "Text analysis and information retrieval of text data," 2016 International Conference on Wireless Communications,Signal Processing and Networking (WiSPNET), Chennai, 2016, pp. 788-792.
[12] P. Shah and N. P. Desai, "A survey of automatic text summarization techniques for Indian and foreign languages," 2016 International Conference on Electrical, Elec- tronics, and Optimization Techniques (ICEEOT), Chennai, 2016, pp. 4598-4601.
[13] S. Krishnan and M. Chen, "Identifying Tweets with Fake News," 2018 IEEE In- ternational Conference on Information Reuse and Integration (IRI), Salt Lake City, UT, 2018, pp. 460-464.
[14] M. U. Islam, F. B. Ashraf, A. I. Abir and M. A. Mottalib "Polarity detection of online news articles based on sentence structure and dynamic dictionary," 2017 20th International Conference of Computer and Information Technology (ICCIT),Dhaka, 2017, pp. 1-5.
[15] Yamada, "Detection of topics from newspaper and its analysis of temporal vari- ations in regions," 2017 Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC), Tainan, 2017, pp. 44-49.