Efficacy of Indian Government Welfare Schemes Using Aspect-Based Sentimental Analysis

Maninder Kaur, Akshay Girdhar, and Inderjeet Singh

Abstract One of the simplest methods to understand people's thoughts using images or text is commonly given as sentiment analysis. Sentiment analysis is used mostly in products advertisement and promotion depends on the user's opinion. The process is based on the aspect-based sentiment analysis and it is used to understand and find out what someone is speaking about, and likeness and dislikeness. One of the real-world models of the perfect realm of this subject is the huge number of available Indian welfare plans like Swachh Bharat Abhiyan and Jan Dhan Yojna. In this paper, labeled data is used on the basis of polarity. Tweets are preprocessed and unigram features are then extracted. In the initial steps, tokenization process, stop word removal process, and stemming process are performed as preprocessing to remove duplicate data. The unigram features and labels trained by support vector machine (SVM), K-nearest neighbor (KNN), and a combination of SVM, KNN, and random forest as a proposed model are used in the presented work. Implementation of experimental proposed approach demonstrates that better results in accuracy and precision than SVM and KNN.

Keywords Sentiment analysis · Aspect · Support vector machine · K-nearest neighbor

M. Kaur (✉)
Department of Computer Science and Engineering, Guru Nanak Dev Engineering College, Ludhiana, India

A. Girdhar
Department of Information Technology, Guru Nanak Dev Engineering College, Ludhiana, India
e-mail: akshay_girdhar@gndec.ac.in

I. Singh
Information Technology, Govt. Polytechnic College, Bathinda, India

A. Pasumpon Pandian et al. (eds.), *Computer Networks, Big Data and IoT*, Lecture Notes on Data Engineering and Communications Technologies 66,
https://doi.org/10.1007/978-981-16-0965-7_74

1 Introduction

In natural language processing, sentiment analysis is considered as most significant tool because it opens up numerous possibilities to understand people's sentiments on different topics. The purpose of an aspect-based sentiment analysis is to detect the features of the particular entity. The positive and negative aspects of a particular topic can be analyzed through aspect-based sentiment analysis [1]. In this research paper, aspect-based sentiment analysis (ASBA) is implemented on the tweets of the government welfare schemes. This type of analysis is mainly domain specific. The government has launched the various welfare schemes for schools, states, as well as center. The given outlines progress in collaboration with center and state governments. The welfare schemes have been mostly introduced to develop the weaker and minority section of the society [2]. Many schemes are launched but in this research work is on Swachh Bharat Abhiyan and Jan Dhan Yojna. These schemes empower every Indian by helping them financially and providing the basic facilities [3].

A. *Two Types of Sentiment Analysis*

Document Level: If the analysis is performed using documents to identify the positive and negative view of the single entity, then it is document level analysis [4].

Comparative: In many cases, users express their views by comparing with the similar product or entity. The main goal is to identify the opinion from the comparative sentence [4].

The research work is structured as follows: Related works are presented in Sect. 2, proposed model is presented in Sect. 3, results and its discussions are presented in Sect. 4, and finally, conclusion is presented in Sect. 5.

2 Related Work

Various approaches proposed by the researchers for sentiment analysis are presented in this section. Shidaganti et al. [5] presented a technique that used data mining along with machine learning as a combination process. Clustering is done by using k-means clustering and hierarchical clustering approach. This approach helps to analyze the data of different organizations which helps in understanding the thoughts and opinions related to the product. Rout et al. [6] presented the way to deal with unstructured data of social media like blogs, Twitter for sentiment and emotion analysis. The supervised and unsupervised approaches are performed on different databases. The unsupervised approach was used for automatic identification of sentiments for the tweets. The sentiment identification is done by using maximum entropy, multinomial Naïve Bayes, and support vector machine classifier. This approach also works well if it will apply on the larger dataset in future. Mumtaz et al. [7] presented an approach which is a combination of lexical-based approach and machine learning. In this proposed work, hybrid approach was used, which gave high accuracy than

the classical lexical method and provides the enhanced redundancy than learning approach. Natural language processing to extract sentiments from texts is performed in the approach. Al-Smadi et al. [8] worked on aspect-based sentiment analysis to review the hotels in Arabic. Long short-term memory (LSTM) neural networks are used in the research model and it was implemented in two levels that are character level and aspect based with random field classifier and polarity classification based. Chen et al. [9] presented a visualization approach called Tag Net which is used for sentiment analysis. This approach combines the improved node-link diagrams with tag clouds to obtain heterogeneous time varying information. Large datasets scalability is improved using the proposed algorithm. Fouad et al. [10] presented a model for Twitter sentiment analysis which describes the tweet is positive or negative by using the concept of machine learning. The presented work uses different methods to label the input in the training phase using different datasets. The classification is also done by using different classifiers to compare their performances. The concept of feature selection and information gain is used in this work. Jones et al. [11] focused on the implication of PMJDY scheme for the development all over India. The main aim of scheme was to make India digitized, even the rural areas are conscious about their bank accounts to obtain government offering benefits directly. Demonetization of currency and all these measures lead toward the progress in the structure of Indian economy. This scheme ensures the better quality of life in the country and helps improve the living status of the people. India is the only nation where 50% of people are in the working age group. Khan et al. [12] focused on the challenges that Twitter information faces, concentrating on order issues, and afterward consider these streams for supposition mining and sentiment analysis. To manage gushing unequal classes, the sliding window Kappa measurement for assessment in time-changing information streams was used. Utilizing this measurement an investigation is performed on Twitter information for utilizing learning calculations for information streams. Greica et al. [13] presented an analytic model which performs aspect classification, separation and polarity classification for analysis. The testing of domain aspect and sentiment classification is performed on the multilingual SemEval dataset. Pham et al. [14] presented the multilayer architecture for representing customer review. It extracts the views for product from the sentiment aspects and sentences related to review. The multilayer architecture represents the different level of sentiments for input text. This model is integrated with neural network for prediction of overall ratings of the product. Rathan et al. [15] worked on the review of mobile phones by using the tweets. In this lexicon-based approach is used for data labeling and this improves the classification process. The support vector machine classifier classifies the tweets with efficient accuracy level. Kim et al. [16] worked on the co-occurrence of data by using supervised and unsupervised methods. A framework is used for processing texts reviews. This work also finds the aspect categories according to review sentences and provides effective outcomes from the F-score. Pannala et al. [17] presented the existing work for the opinion mining that was done on the word level, not on the sentence level. The work was done on the trained dataset. In this author presents the combination of natural language (NL) and machine learning (ML) model to process the dataset which has 1654 aspects in the

training dataset with different category annotations and 845 aspects in the test dataset with different category annotations for analysis.

The performance of software is measured by SVM and logistics regression algorithm. Previously, work was done on the static parameters which reduce the effective learning of tweets according to its label. The nonparametric approach uses number of coefficient parameters those increase the over fitting. In proposed paper, the hyperparameters are used which reduces the adaptive features of learning. The analysis is on the combined performance of parametric (SVM), nonparametric (KNN), and the random forest is used.

3 Methodology

The data is collected for the experiment from Twitter and stored in a database for preprocessing. After the preprocessing on the feature label, the processed data is learned by KNN, SVM, and proposed model which is the combination of (KNN, SVM, random forest) these. By these respective models, tweets are predicted and classified, respectively. The below-given section describes the proposed methodology of the model and the techniques used in this work in detail. The pictorial representation of the proposed model is presented in Fig. 1.

Step 1: Collection of data
The proposed model used the data collected from the Twitter, regarding government welfare schemes as input. People get easy access to financial and banking services due to Jan Dhan Yojna of government welfare schemes and the main aim of Swachh Bharat Abhiyan is to keep India clean [18]; tweets regarding both the scheme have been used to determine the public review based on aspect-based sentiment analysis. The data that retrieved from the social media is in unstructured form due to intensive information.

Step 2: Data Storing and fetching
The retrieved tweets are put in storage as .csv files, and it is fetched using a Python tool PyCharm [19]. Around 3000 tweets are stored for the purpose of training and testing the datasets. Data mining algorithm (SVM, KNN, and hybrid SVM-KNN) is utilized to train and test the fetched tweets.

Step 3: Preprocessing
The redundant and noise contents are removed in preprocessing step which makes the data suitable for training process [20].

Data cleaning is performed based on the following steps.

(i) All the uppercase is converted to lowercase.
(ii) Remove all the Internet slangs from the data.
(iii) Removing all the stopping words from the list.

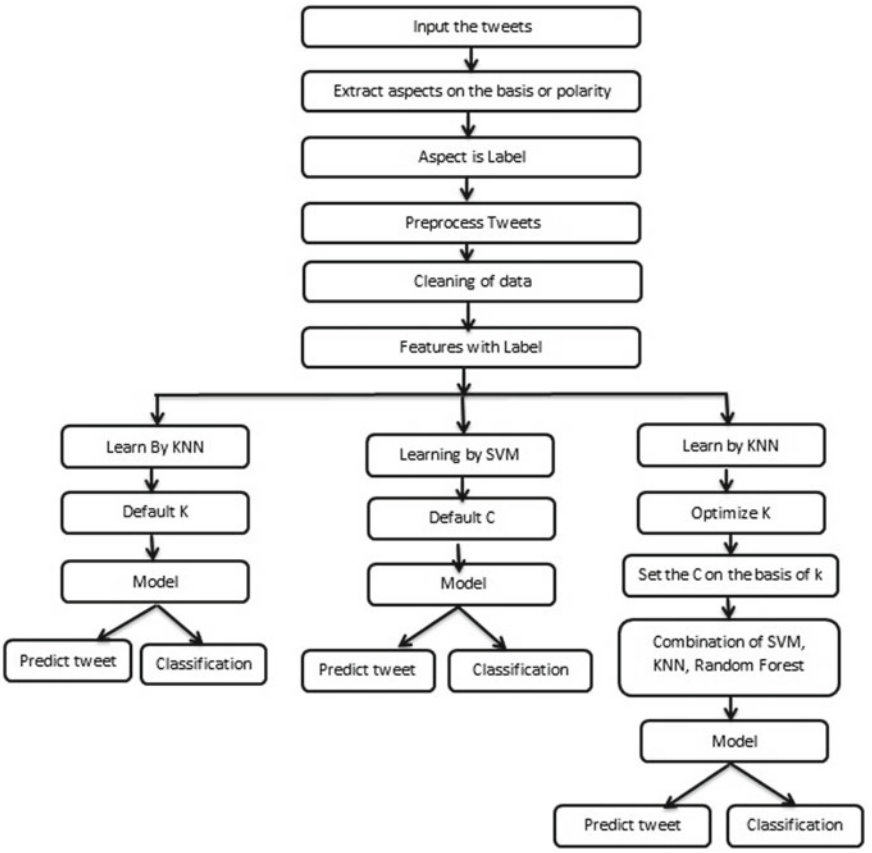

Fig. 1 Proposed framework process flow

(iv) Eliminating all the additional white spaces.
(v) Compress the duplicate words.
(vi) All the hashtags are removed but the hashtags texts are selected.

Step 4: Classification using data mining techniques

Classification is performed using data mining techniques to categorize the data into various aspects such as .

First aspect: increase fund/decrease fund.
Second aspect: improvement in growth/not growth/growth.
Third aspect: goes really fast/works/hard fix.
Fourth aspect: incredible/good/not good.
On the basis of these parameters, the data is trained and tested.

Machine learning models are trained using the training datasets. In the proposed framework, training data is implemented for classification, then testing dataset is prepared that is not before used in the proposed model for training.

Step 5: Optimize the classification results
The classification results are need to be checked to confirm whether the learning models training datasets follow the defined rules or not. This process is performed to obtain accurate and error-free results. Python is used to train and test the proposed KNN-SVM random forest model. The tweets nature is predicted as optimal value based on the classification results.

4 Results and Discussion

The proposed model classification performance is evaluated and compared with conventional approaches to validate the better performance. The experiment result validation is done by using fivefold and tenfold validation approach. The data is divided into k subsets in cross validation which has equal size and the training process and testing process are repeated for k-times. Every time one group of subset will be used for testing of data and other k-1 subsets of data will be treated as training data. The result analysis is done on SVM, KNN, and the combination of proposed model (SVM, KNN and random forest algorithm). The parameters used in this for analysis process are accuracy, precision, recall, and F-measure.

The classification accuracy of all the three models is depicted in Fig. 2. It is observed that the proposed shows the maximum accuracy in both fold testing process. KNN obtains the least accuracy performance of 49.23%, 51.23%, respectively, for both processes.

The precision analysis for proposed model and conventional SVM and KNN model is depicted in Fig. 3. The proposed algorithm shows the maximum precision and KNN obtains minimum precision 50.23, 51.23 in fivefold and tenfold validation process, respectively (Fig. 4).

The overall performance of all the classifiers is presented in figure. It is observed that the proposed model (KNN-SVM-random forest) shows the optimum, enhanced results with 82% in comparison with the SVM and KNN.

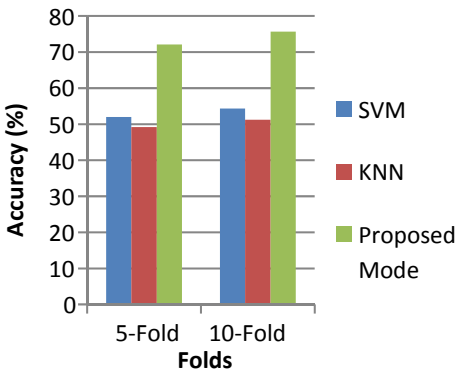

Fig. 2 Accuracy comparison

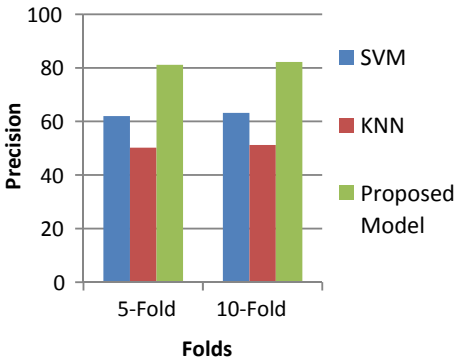

Fig. 3 Precision analysis

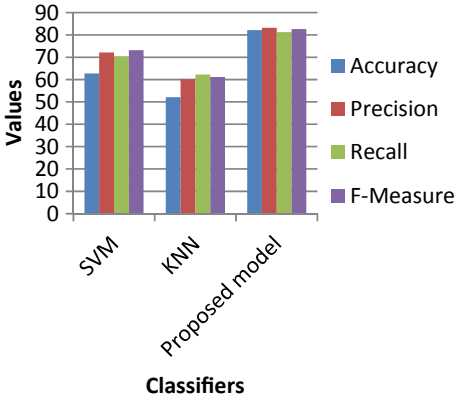

Fig. 4 Overall performance comparison

5 Conclusion

A hybrid classifier is presented in this research work as sentiment analysis model. The proposed hybrid model obtains better results by utilizing part-of-speech (POS) tags and word dependencies for aspect-based sentiment analysis. Efficacy of government welfare schemes using the proposed model is computed 82% under the restricted environment. In future work, improvement in the accuracy will be considered by reducing the feature sparsely by divergence and optimization approaches, improve the classifier by deep learning approaches. Additionally, other tasks that are Aspect Based Sentiment analysis also tested on the applicability on best of learning and concerns with the integration of POS tags, word dependencies, and possibly other NLP tools.

References

1. Perikos I, Hatzily geroudis I (2017) Aspect based sentiment analysis in social media with classifier ensembles. In: 2017 IEEE/ACIS 16th international conference on computer and information science (ICIS), Wuhan, China, pp 273–278
2. Tiwari SK (2014) To study awareness of a national mission: Swachh Bharat: Swachh Vidyalaya in the middle school student of private and public schools. Paripex-Indian J Res 3(12):23–24
3. Barhate GH, Jagtap VR (2014) Pradhan Mantri Jan Dhan Yojana: national mission on financial inclusion. Indian J Appl Res 4(12):340–342
4. Lin Y, Zhang J, Wang X, Zhou A (2012) An information theoretic approach to sentiment polarity classification. In: Proceedings of the 2nd joint WICOW/AIRWeb workshop on web quality, Lyon,France, 2012, pp 35–40
5. Shidaganti G, Hulkund RG, Prakash S (2017) Analysis and exploitation of Twitter data using machine learning techniques. In International proceedings on advances in soft computing, intelligent systems and applications, pp 135–146
6. Rout JK et al (2018) A model for sentiment and emotion analysis of unstructured social media text. Electr Commerce Res 18(1):181–199
7. Mumtaz D, Ahuja B (2017) A lexical and machine learning-based hybrid system for sentiment analysis. In: Studies in computational intelligence. Springer, Singapore (Chap. 11, pp 165–175)
8. Al-Smadi M, Talafha B, Al-Ayyoub M, Yaser J (2018) Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews. Int J Mach Learn Cybern 1–13
9. Chen Y (2018) TagNet: toward tag-based sentiment analysis of large social media data. In: 2018 IEEE Pacific visualization symposium (PacificVis), Kobe, Japan, pp 190–194
10. Fouad Mohammed M, Gharib Tarek F, Mashat Abdulfattah S (2018) Efficient Twitter sentiment analysis system with feature selection and classifier ensemble. In: International conference on advanced machine learning technologies and applications, vol 723, pp 516–527
11. Mary Jones T, DivyaSri S, Bavani G (2017) A study on the implications of Pradhan Manthri Jan Dhan Yojana on the growth of indian economy. IRA-Int J Manag Soc Sci 6(3):461–466
12. Khan FH, Bashir S, Qamar U (2014) TOM: Twitter opinion mining framework using hybrid classification scheme. Decis Support Syst 57:245–257
13. García-Pablos A, Cuadros M, Rigau G (2018) W2VLDA: almost unsupervised system for aspect based sentiment analysis. Expert Syst Appl 91:127–137
14. Pham D-H, Le AX (2018) Learning multiple layers of knowledge representation for aspect based sentiment analysis. Data Knowl Eng 114:26–39
15. Rathan M, Hulipalled VR, Venugopal KR, Patnaik LM (2018) Consumer insight mining: aspect based Twitter opinion mining of mobile phone reviews. Appl Soft Comput 68:765–773
16. Schouten K, van der Wejide O, Frasincar F, Dekker R (2018) Supervised and unsupervised aspect category detection for sentiment analysis with co-occurrence data. IEEE Trans Cybern 48(4):1263–1275
17. Pannala NU, Nawarathna CP, Jayakody JTK, Rupasinghe L, Krishnadeva K (2017) Supervised learning based approach to aspect based sentiment analysis. In: 2016 IEEE international conference on computer and information technology (CIT), Nadi, Fiji, pp 662–666
18. Thakkar P (2015) Swachh Bharat [CLEAN INDIA] Mission—an analytical study. Renew Res J 3(2):168–173
19. Massiris MM, Dennehy BR, Delrieux CA, Thomsen FSL (2017) Python implementation of local intervoxel-texture operators in neuroimaging using Anaconda and 3D Slicer environments. In: 2017 XLIII Latin American Computer Conference (CLEI), Cordoba, Argentina, pp 1–3
20. Dwivedi SK, Rawat B (2016) A review paper on data preprocessing: a critical phase in web usage mining process. In: 2015 international conference on green computing and Internet of Things (ICGCIoT), Noida, India, pp 506–510