International Journal of Engineering & Technology, 7 (3) (2018) 1372-1375

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i3.12656
Research paper

KNN classifier based approach for multi-class sentiment analysis of twitter data

Soudamini Hota [1], Sudhir Pathak [1] *

[1] Chandigarh University, Mohali, India
**Corresponding author E-mail: pathak.cse@cumail.in*

Abstract

‘Sentiment’ literally means ‘Emotions’. Sentiment analysis, synonymous to opinion mining, is a type of data mining that refers to the analy-sis of data obtained from microblogging sites, social media updates, online news reports, user reviews etc., in order to study the sentiments of the people towards an event, organization, product, brand, person etc. In this work, sentiment classification is done into multiple classes. The proposed methodology based on KNN classification algorithm shows an improvement over one of the existing methodologies which is based on SVM classification algorithm. The data used for analysis has been taken from Twitter, this being the most popular microblogging site. The source data has been extracted from Twitter using Python’s Tweepy. N-Gram modeling technique has been used for feature extraction and the supervised machine learning algorithm k-nearest neighbor has been used for sentiment classification. The performance of proposed and existing techniques is compared in terms of accuracy, precision and recall. It is analyzed and concluded that the proposed technique performs better in terms of all the standard evaluation parameters.

Keywords: *KNN; N-Gram; SVM; Tweedy; Twitter*

1. Introduction

Social networking platforms have been gaining huge popularity amongst people in the recent years. These online sites are being widely used by people to express their emotions, beliefs as well as opinions towards any entity ranging from product, person, event and so on. These networking sites provide a platform for users to post their feedback and reviews and the data generated therein are being harnessed by business enterprises in order to get an insight into how well their products and services are faring in the market. This knowledge helps business analysts and managers in better decision making. Apart from business enterprises, sentiment analysis of user comments is of immense use for buyers too. For instance, if someone wants to either buy a product or access any service, generally the initial step would be to go through online reviews and generate a discussion regarding it on the social media before taking any decision. However it is not possible for a user to analyze all the reviews considering the massive amount of user reviews and comments available on online platforms. Hence several sentiment analysis techniques have been proposed in order to automate this analysis process [1], [2]. Through these techniques a user would be able to know about the positive as well as negative views that the other users have regarding a product. Thus the user gets a clear view about the products and services, and can assess whether it is as per the requirements [3]. In order to gather, process, and analyze the factual data, text information retrieval techniques are used. A textual comment consists of both objective and subjective components. The analysis and identification of the subjective components that express opinions, sentiments, and attitudes is of immense help to manufacturers and service providers in fine tuning their business strategies. Likewise, it helps customers and clients to take better decisions while they buy a product or avail a service [4], [5]. Lately, automatic sentiment analysis as well as opinion mining has been the most popular topic for study and research [6], [7]. A significant proportion of customers and clients who utilize a product or service generate a huge amount of data in the form of comments, feedbacks and reviews that express their opinions. However, developing applications for analyzing such data involves several challenges, considering the enormous size of such data available and the structure of the data [8], [9]. The presence of informal words, slang words, and abbreviations make it difficult to identify and classify sentiments. Also for the same word, multiple meanings with completely different sentiment polarities can be derived; this task is known as “Handling Polysemy”. The meaning and polarity of slang words are identified by referring dictionary of slangs, informal words are replaced by their synonyms, and abbreviations are replaced by their expanded forms.

Sentiment classification is either binary or ternary within most of the proposed approaches i.e. classification of text is done into either, “positive” and “negative” or “positive”, “negative” and “neutral”. Twitter is a microblogging site where users can post real time messages known as tweets [10]. The unique properties of tweets pose new challenges in sentiment classification methods. Certain important characteristics of tweets are discussed further. Initially the maximum permissible length of tweets was 140 characters which was doubled in November, 2017, for all languages except Japanese, Korean and Chinese. Users include several acronyms, misspellings, cyber slangs, emoticons and other characters with special meanings in order to make their messages quick and short. Another interesting characteristic is that Twitter permits tweets on a wide range of topics instead of focusing on any particular topic or theme. Tweets can be updated in real time since their size is limited which also makes them less time consuming [11]. This makes it better than blogs for sentiment analysis as blogs are longer in nature and more time consuming which is why the updating of blogs is done at longer intervals. There are few basic terminologies used within Twitter applications. The representations of facial expressions that are generated using different combinations of punctuations and letters are known as emoticons. These pictorial representations effectively convey the mood of the user. Another important terminology relates to the facility to mention target usernames along with "@" symbol. If the users are mentioned in this manner, they receive alerts automatically. Hashtag is yet another significant feature of tweets. In order to stamp topics, "#", known as hash-tag is used. Through hash-tags, it is possible for larger number of audience to view the tweets. All these properties of tweets are taken into consideration along with the regular textual features during feature selection [12-15] [16-17]. Twitter is the most popular microblogging site and hence generates enormous amounts of data suitable for opinion mining as compared to any other social media.

2. Literature review

Zhao Jianqiang, et al. [18] has proposed an improvement over the conventional sentimental analysis approaches which focus on lexical analysis of every unit of the tweets like words, exclamation marks, emoticons etc. The proposed approach is based on unsupervised learning method that forms word embeddings with reference to huge twitter corpora by considering the latent semantic relationships of words based on contextual meaning, as well as co-occurring statistical properties between words in the tweets. In order to generate feature set from the tweets, the word embeddings are integrated with n-gram features and the sentiment polarity score features of the words. To train and label the sentiment classifier, the feature set is incorporated into a deep convolution neural network. The performance scores of the proposed model is compared with that of the base model on five Twitter datasets, wherein the proposed model shows better performance in terms of accuracy and F1-measure.

K Lavanya, et al. [19] proposed a classification mechanism based on a training method that gets adapted to the topic of the tweets. This mechanism has been proposed to address the problem of handling vast diversity of topics in Twitter. This makes the classification algorithm dynamic in nature. In this approach, the non-textual features of tweets are also used for training the classification algorithm. The proposed methodology can be applied on static data belonging to completely different topics as well as on dynamic data for a particular timeline while they are streaming in. The classification algorithm categorizes tweets into three primary classes namely positive, neutral and negative. These three class labels can be further extended to form five class labels namely neutral, positive, very positive, negative and very negative. The proposed approach shows an improvement in performance with respect to recall, precision and F-score.

Chintan Dedhia, et al. [20] proposed an add-on method to improve the strength of SVM classifier for sentiment classification which is one of the commonly used machine learning algorithms for sentiment analysis and classification. This ensemble model is designed by integrating SVM algorithm as the base classifier coupled with Adaboost algorithm for Ensemble boosting. The proposed algorithm makes use of the structured details associated with the tweets like retweets, followers of tweets, tags inside the tweets along with all other essential characteristics of tweets in order to find the relationship information between tweets. This helps in studying the properties of social interactions. The proposed ensemble model is used to classify Twitter data into positive and negative labels. The proposed approach has been compared with the baseline SVM algorithm and it demonstrates better performance with respect to precision, recall and F-score.

Yeqing Yan, et al. [21] introduced two simple but potent ensemble classifiers to perform sentiment classification of twitter data. The two ensemble models are developed using the classifiers Naïve Bayes with Mallet's MaxEnt, and SentiStrength with Pattern of Textblob. These model overcomes the problem that arises in training classifiers properly when there isn't enough training data; i.e., in order to categorize the tweets related to one product, the tweets related to other similar products can be used to learn the classifier. Both the models have proved to be efficient by demonstrating high accuracy in sentiment classification of twelve different datasets. Both the ensemble classifiers can be executed in parallel and are capable of handling large sets of Twitter data.

Paramita Ray et al. [22] proposed a R-based framework built on lexicon based approach for sentiment analysis and categorization of product reviews, which can be helpful in better decision making related to products and services. The preprocessing of input tweets primarily includes replacing acronyms with their expanded forms, replacing emoticons with the words describing the associated sentiments, removal of stop words and handling of negations. The proposed methodology analyzes text at both document level as well as aspect level. The author aims at developing in future a hybrid working model based on machine learning techniques for sentiment analysis and classification.

Ranjan Satapathy, et al. [23] mentioned in his paper that the concept of microtext which became prevalent due to the rising use of Web 2.0 technology posed challenges to standard natural language processing tools as these tools are designed to handle well-formed text. Normalization of microtext helps them overcome these challenges. Hence the author has proposed an approach based on phonetics to transform micro text into plain text in English. Demonstration results show that the similarity index between the tweets normalized by the proposed technique and the tweets normalized by human annotators is equal to or more than 0.8 for 85.31% of the tweets. It is also found that normalization of tweets improves the accuracy in polarity detection by more than 4%.

3. Proposed methodology

This research work is based on sentiment identification and classification of Twitter data. In the existing system, a sentiment analysis and classification tool has been developed based on SVM classification algorithm, which can classify Twitter data into seven classes. In this research work, KNN classification algorithm has been used to model a classifier for sentiment analysis and classification of Twitter data into seven classes. The proposed methodology is depicted in Fig.1. The performance of both the classifiers has been compared in terms of accuracy, precision and recall.

3.1. SVM classifier

Support Vector Machine SVM is an algorithm that can be used for linear and non-linear classification, as well as linear and non-linear regression. A binary classifier is the initial form of SVM classifier in which the learned function classifies the data into positive and negative classes [24]. A multiclass classifier can be modeled by the integration of multiple binary classifiers using pair-wise coupling method. The maximum margin hyperplane that linearly separates n-dimensional data points X={$x_{1,}$ x2, x3… x_n} with attribute weights W= {w_1, w2, w3$_{…}$ w_n} into two classes

Yi= {+1, -1} is defined by the margins

$$H_1: w_0+\sum^n_{k=1} w_k x_k \geq +1 \text{ for } y_i = +1 \quad (1)$$

$$H_{2:} w_0+\sum^n_{k=1} w_k x_k \leq -1 \text{ for } y_i = -1 \quad (2)$$

Where w_0 is a scalar known as bias

Combining the two inequalities of equations (1) and (2)

$$Yi\,(w_0 + \sum^n_{k=1} w_a x_{)} \geq 1, \text{ for all values of i} \quad (3)$$

This is a convex quadratic optimization problem which is solved to obtain the support vectors and maximum margin hyperplane. Based on Langrangian formulation, the maximum margin hyperplane is defined by

$$D\,(X^T) = \sum\nolimits_{i=1}^{l} y_i\alpha_i X_i X^T + b \qquad (4)$$

Where y_i is the class label of support vector X_i, l is the number of support vectors, X^T is a test tuple, α_i and b are numeric parameters determined automatically by the optimization of SVM algorithm.

For supporting non-linear classification, non-linear mapping is done to transform the original data in the input space to a higher dimensional space. By avoiding exact formulation of the mapping function, the kernel trick is used through which the curse of dimensionality is generated. Thus the non-linear classification within original space is equated to linear classification in the new space. Within the new dimensions, SVM classifier searches for maximum margin hyperplane that separates the data points of one class from another.

3.2 KNN classifier

K-Nearest-Neighbors (KNN) is a non-parametric supervised classification algorithm, which is simple yet effective in many cases. The KNN classifier is considered as the most popular classifier for pattern recognition due to its effective performance with efficient results and its simplicity. It is widely used in the field of pattern recognition, machine learning, text categorization, data mining, object recognition and many more [25]. KNN algorithm classifies by analogy i.e. by comparing the unknown data point with the training data points to which it is similar. Similarity is measured by Euclidean distance. The attribute values are normalized to prevent attributes with larger ranges from outweighing attributes with smaller ranges. In KNN classification, the unknown pattern is assigned the most predominant class amongst the classes of its nearest neighbors. In case there is a tie between two classes for the pattern, the class that has minimum average distance to the unknown pattern is assigned. Through the combination of a number of local distance functions based on individual attributes, a global distance function dist can be calculated. As given in equation (5), the simplest way is to sum up the values:

$$\text{dist.}\ (X^T, X) = \sum\nolimits_{i=1}^{n} \text{distA}_i\,(X^T{\cdot}A_i,\ X{\cdot}A_i) \qquad (5)$$

Where X^T is the test tuple, X is a nearest neighbor, and A_i (i=one to n) represents the attributes of the data points.

The weighted sum of local distances is known as global distance. The attributes A_i can be assigned specific weights w_i to depict their level of importance in deciding the appropriate classes for the samples. The weights usually range between 0-1. Irrelevant attributes are assigned a weight 0. Thus, equation (4) can be modified and written as equation (6):

$$\text{dist}(X^T,X) = \sum\nolimits_{i=1}^{n} w_i \times \text{distA}_i(X^T{\cdot}A_i\ ,\ X{\cdot}A_i) \qquad (6)$$

The average weighted distance is given by equation (7):

$$\text{avgdist}(X^T,X) = \sum\nolimits_{i=1}^{n} w_i \times \text{distA}_i(X^T{\cdot}A_i\ ,\ X{\cdot}A_i) \qquad (7)$$

$$\sum\nolimits_{i=1}^{n} w_i$$

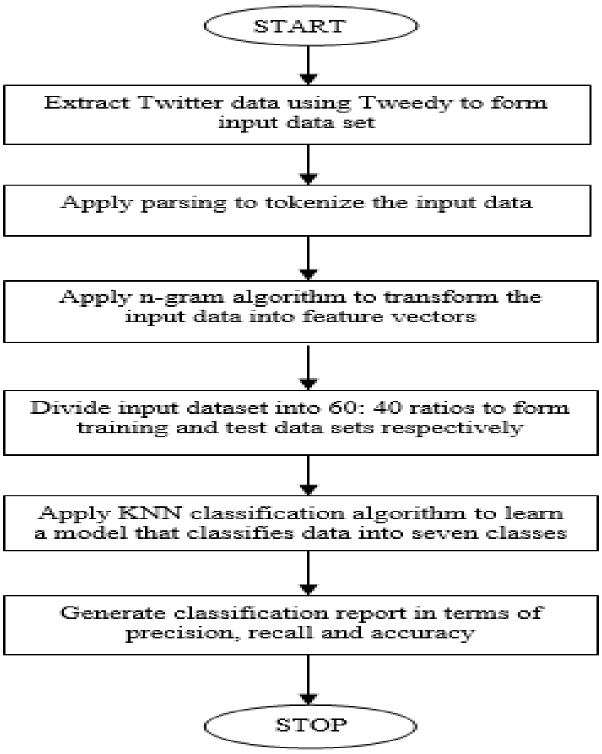
START
Extract Twitter data using Tweedy to form input data set
Apply parsing to tokenize the input data
Apply n-gram algorithm to transform the input data into feature vectors
Divide input dataset into 60: 40 ratios to form training and test data sets respectively
Apply KNN classification algorithm to learn a model that classifies data into seven classes
Generate classification report in terms of precision, recall and accuracy
STOP

Fig. 1: Proposed Methodology.

4. Results and analysis

The performance of the proposed system is analyzed on the basis of various performance analysis metrics namely precision, recall and accuracy. The performance of the proposed system is compared with that of the existing system in which SVM classifier is used for the classification of twitter data into seven classes. The formulas that define precision, recall and accuracy are given by equations (8), (9) and (10) respectively.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive + False Positive}} \qquad (8)$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive + False Negative}} \qquad (9)$$

$$\text{Accuracy} = \frac{\text{No. of tweets correctly classified}}{\text{Total no. of tweets}} \qquad (10)$$

The precision of the proposed system is approx. 82 percent, while the precision of the existing system is approx. 79 percent. The recall of the proposed system is 81.5 percent whereas recall value of the existing system is up to 78 percent. The accuracy achieved by the proposed system is up to 86 percent whereas accuracy of the existing system is approx. 81 percent. The performance evaluation of the proposed and existing systems is shown in Table 1 and Figure 2.

Table 1: Comparison of Performance Measures

Performance Metrics	Existing System	Proposed System
Precision	79 percent	82 percent
Recall	78 percent	81.5 percent
Accuracy	81 percent	86 percent

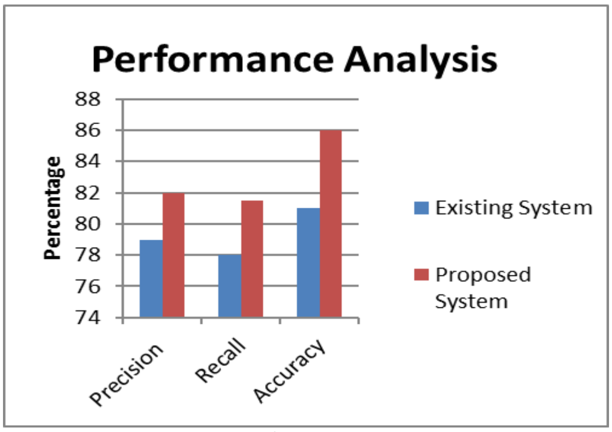

Fig. 2: Performance Analysis.

5. Conclusion

In this work, it is concluded that sentiment analysis, a technique to analyze and classify the sentiments conveyed by user statements, can be carried out using several methodologies. One of the existing methodologies is based on employing SVM algorithm to train the classifier that distinguishes twitter data into multiple classes. In this research work, the aforementioned methodology is improved further by using KNN algorithm to train the classifier. The performance of KNN algorithm is improved significantly with the right choice of the parameter k, the use of an appropriate distance metric, incorporating attribute weighting and pruning of noisy data points. The existing and proposed techniques are implemented in Python and simulation results show that the proposed approach scores over the existing approach in terms of precision, recall and accuracy. The proposed approach can be improved further by employing distance weighted KNN algorithm that involves associating weights with the nearest neighbors based on their proximity to the data point; the nearest the neighbor the greater the weight. This would enable fine tuning of the classification process. Also, for extensive data sets, the computational cost can be reduced by indexing the training data points using tree structures like KD Tree and Ball Tree.

References

[1] Meesala Shobha Rani, Sumathy S, “Perspectives of the performance metrics in Lexicon and Hybrid based approaches: a review”, IJET, Vol. 6, No 4, 2017.

[2] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, New Avenues in Opinion Mining and Sentiment Analysis, IEEE Intelligent Systems, 28(2), 2013, pp 15-21. https://doi.org/10.1109/MIS.2013.30.

[3] Sasikumar.A.N, “Sentimental Analysis of Social Networking Sites for Categorization of Product Reviews”, Internation Journal of Pure and Applied Mathematics, Vol. 117, 2017, pp. 87 – 92.

[4] J. Mannar Mannan, J, Jayavel, “An adaptive sentimental analysis using ontology for retail market”, IJET, Vol.7, No 1.2, 2018.

[5] V. Uma. Ramya, K. Thirupathi Rao, “Sentiment Analysis of movie review using Machine Learning techniques”, IJET, Vol.7 (2.7), 2018.

[6] Thelwall, M., Buckley, K., & Paltoglou, G., "Sentiment strength detection for the social web", J.American Society for Information Science and Technology, Vol.63 No1, 2016, pp.163–173. https://doi.org/10.1002/asi.21662.

[7] Paltoglou, G, & Thelwall, M. "Twitter, MySpace, Digg Unsupervised Sentiment Analysis in Social Media", ACM Transactions on Intelligent Systems and Technology, Vol.3 No 4, 2012, pp.1-19. https://doi.org/10.1145/2337542.2337551.

[8] Wafa Zubair Al-Dyani, Adnan Hussein Yahya, Farzana Kabir Ahmad, “Challenges of Event Detection from social media streams”, IJET, Vol.7 (2.15), 2018.

[9] Socher, R., Perelygin, A., Y.Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C., “Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank”, In the Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Washington, USA, 18-21 October 2013, Vol. 1631, pp. 1642.

[10] Alexander Pak & Patrick Paroubek. "Twitter as a corpus for sentiment analysis and opinion mining", In Proc. LREC, Vol.10, 2010, pp.1320-1326.

[11] Ankur Goel, Jyoti Gautam, Sitesh Kumar, “Real time sentiment analysis of tweets using Naive Bayes”, 2nd International Conference on Next Generation Computing Technologies (NGCT), 2016.

[12] P. Lalitha Kumari, Ch Sathyanarayana, “A novel cluster based feature selection and document classification model on high dimension tree data”, IJET, Vol.7,No1.1, 2018.

[13] Tang, D., FuruWei, Yang, N., Zhou, M., Liu, T., & Qin, B., “Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification”, In the Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics Baltimore, Maryland, USA, June 23-25 2014, pp. 1555-1565.

[14] Wiraj Udara Wickramaarachchi, R. K. A. R. Kariapper, “An Approach to Get Overall Emotion from Comment Text towards a Certain Image Uploaded to Social Network Using Latent Semantic Analysis”, 2nd International Conference on Image, Vision and Computing, 2017.

[15] Ratna Sathappan, Tholu Sai Indira, A. Meenapriyadarsini, “Smart Recommendation System for off-the shelf medicines”, IJET, Vol.6, No. 2.24, 2017.

[16] Tharindu Weerasooriya, Nandula Perera, S.R. Liyanage, “A method to extract essential keywords from tweet using NLP”, 16th International Conference on Advances in ICT for Emerging Regions (ICTer), 2016.

[17] Prasanna Moorthi N, Mathivanan V, “An improved Wrapper based feature selection for feature mining”, IJET, Vol. 7 (1.3), 2018.

[18] Zhao Jianqiang, Gui Xiaolin, “Deep Convolution Neural Networks for Twitter Sentiment Analysis”, IEEE, 2017.

[19] K Lavanya, C Deisy, “Twitter Sentiment Analysis Using Multi-Class SVM”, International Conference on Intelligent Computing and Control (I2C2'17), 2017.

[20] Chintan Dedhia, Mrs Jyoti Ramteke, “Ensemble model for Twitter Sentiment Analysis”, International Conference on Inventive Systems and Control (ICISC), 2017. https://doi.org/10.1109/ICISC.2017.8068711.

[21] Yeqing Yan, Hui Yang, Hui-ming Wang, “Two Simple and Effective Ensemble Classifiers for Twitter Sentiment Analysis”, Computing Conference, 2017.

[22] Paramita Ray and Amlan Chakrabarti, “Twitter Sentiment Analysis for Product Review Using Lexicon Method”, International Conference on Data Management, Analytics and Innovation (ICDMAI), 2017. https://doi.org/10.1109/ICDMAI.2017.8073512.

[23] Ranjan Satapathy, Claudia Guerreiro, Iti Chaturvedi, Erik Cambria, “Phonetic-Based Microtext Normalization for Twitter Sentiment Analysis”, IEEE International Conference on Data Mining Workshops, 2017.

[24] Shweta Rana, Archana Singh, “Comparative analysis of sentiment orientation using SVM and Naive Bayes techniques”, second International Conference on Next Generation Computing Technologies (NGCT), 2016.

[25] Pedro, J., Garcia-Laencina, Jose-Luis Sancho-Gomez, Anibal, R., Figueiras-Vidal, and Michel Verleysen, “K nearest neighbours with mutual information for simultaneous classification and missing data imputation”, 2009, 1483–1493.