Mehran University Research Journal of Engineering & Technology
Vol. 38, No. 2, 463-470 April 2019
p-ISSN: 0254-7821, e-ISSN: 2413-7219
DOI: 10.22581/muet1982.1902.20

Sentiment Analysis for Roman Urdu

AYESHA RAFIQUE*, MUHAMMAD KAMRAN MALIK*, ZUBAIR NAWAZ* FAISAL BUKHARI*, AND AKHTAR HUSSAIN JALBANI**

RECEIVED ON 16.04.2018 ACCEPTED ON 25.05.2018

ABSTRACT

The majority of online comments/opinions are written in text-free format. Sentiment Analysis can be used as a measure to express the polarity (positive/negative) of comments/opinions. These comments/ opinions can be in different languages i.e. English, Urdu, Roman Urdu, Hindi, Arabic etc. Mostly, people have worked on the sentiment analysis of the English language. Very limited research work has been done in Urdu or Roman Urdu languages. Whereas, Hindi/Urdu is the third largest language in the world. In this paper, we focus on the sentiment analysis of comments/opinions in Roman Urdu. There is no publicly available Roman Urdu public opinion dataset. We prepare a dataset by taking comments/opinions of people in Roman Urdu from different websites. Three supervised machine learning algorithms namely NB (Naive Bayes), LRSGD (Logistic Regression with Stochastic Gradient Descent) and SVM (Support Vector Machine) have been applied on this dataset. From results of experiments, it can be concluded that SVM performs better than NB and LRSGD in terms of accuracy. In case of SVM, an accuracy of 87.22% is achieved.

Key Words: Roman Urdu, Sentiment Analysis, Feature Selection, Machine Learning Classifiers.

1. INTRODUCTION

Sentiment analysis is an emerging field of Text Mining and NLP (Natural Language Processing). It uses machine learning and NLP techniques in order to reveal the mood, feelings, beliefs, emotions or attitude of the people based on the text they write. The text could be opinion, review, tweet or something whose sentiment is to be evaluated. The sentiment could be negative, positive or neutral. Sentiment analysis of the text has been done using following approaches:

- List based approach (SentiWordNet) SentiWordNet is a lexical resource for opinion mining. url: http://sentiwordnet.isti.cnr.it/
- Machine learning approach
- Semi-supervised approach
- Combination of above

Due to extensive use of internet, people use social media to share their thoughts, opinions and feelings to the public [1].

Authors E-Mail: (znawaz@pucit.edu.pk, kamran.malik@pucit.edu.pk, faisal.bukhari@pucit.edu.pk, jalbaniakhtar@gmail.com)
* Department of Information Technology, University College of Information Technology, University of the Punjab, Lahore, Pakistan.
** Department of Information Technology, Quaid-e-Awam University of Engineering, Science & Technology, Nawabshah, Pakistan.

People not only obtain information from internet but also vigorously generate content in the form of opinions and comments on the social media. People can also see these newly generated opinions and comments. According to a survey, people reading these opinions and comments are influenced from these opinions and comments [2-3]. For example, we intend to visit a restaurant most likely for which most people have given a high rating online. Similarly, one can decide to watch a movie or a series based on the positive reviews given by others. Usually, people post their comments/opinions about a particular product, service, movie or series in an unstructured form. They prefer to write comments in their native language. Only a few people can understand these comments who speaks that language. So, these comments are not very useful for all the users [4]. What if we make these comments useful for non-native users? The main goal of this paper is to help the non-native Urdu users to know the polarity of others opinion about any product or service from the comments posted in Roman Urdu.

In this paper, an attempt has been made to perform sentiment analysis on comments/opinions in Roman Urdu. First of all, opinion/comments data is collected from various online sources. Then, it is preprocessed by removing stop words, punctuation marks and numerical characters. Three supervised machine learning techniques NB, LRSGD and SVM have been applied on the data and different performance measures have been studied. Every algorithm has its own advantages and disadvantages in terms of model complexity and accuracy.

The contributions of the paper are:

(1) There is no publicly available Roman Urdu opinion; therefore, we have prepared a dataset by taking the comments/opinions of people in Roman Urdu from different websites. Besides that, we have also tagged them in to positive, negative and neutral sentiments.

(2) We have achieved an accuracy of 87.22%, when SVM is used with Unigram + Bigram + TF-IDF as feature set.

The rest of the paper is organized as follows. Related work has been explained in Section II. Section III presents the methodology to analyze sentiment analysis. Experimental setup, results and discussions are briefed in Section IV. Finally, the paper is concluded in Section V.

2. RELATED WORK

A lot of work has been done on the sentiment analysis of content in English and other developed languages. A very little research is done on the sentiment analysis of the content in Roman Urdu/Hindi, the third largest language in world [5]. In this section, we first present some relevant work in English language. At the end of this section, we also discuss some work done on the sentiment analysis of the content in Urdu. Tripathy et. al. [6] worked on sentiment analysis of movie reviews in English language using Machine Learning Techniques. They preprocess data by removing top words, punctuation characters and numerical characters. A numerical matrix, TF-IDF (Term Frequency-Inverse Document Frequency) were generated using labeled polarity movie dataset where rows represent reviews and columns represent features. Machine learning algorithms (NB, SVM) used this matrix as input in order to train the model. They tested this model and studied different performance measures. Then, they critically examined their results on the basis of comparison with existing literature and conclude that their results are better than existing literature. They achieved 94% accuracy for SVM and observed that SVM classifier performed better than every other classifier in predicting the sentiment of movie reviews.

Shaziya et. al. [7] worked on text classification for sentiment analysis on movie review dataset. Their dataset contained 2000 reviews, 1000 positive and 1000 negative. They used WEKA (Waikato Environment for Knowledge Analysis) for experiments and applied different filters on the dataset to find optimal feature set. Later, they performed experiments by applying classification algorithms like NB and SVM on the dataset and found that in movie review sentiment analysis NB performed better than SVM in terms of accuracy. They achieved 85.1% accuracy in case of NB.

Kalaivani et. al. [8] used three machine learning algorithms (NB, KNN and SVM) for sentiment analysis on movie reviews in English language. They apply applied 3-fold cross validation on 1000 positive and 1000 negative reviews. For training dataset, 2-folds were used and one fold used for testing dataset. Training dataset were further divided into ten non non-overlapping sets of different sizes. Then, they examined the performance of all three algorithms on the basis of results of 10 experiments that were performed on 10 train-test sets. They concluded that NB and SVM perform better than KNN but SVM gives best results and achieves more than 80% accuracy when size of training dataset is appropriately large i.e. 800-1000 reviews.

Pang et. al. [9] used machine learning techniques for document level sentiment analysis on movie review dataset. They got results for NB, Maximum Entropy and SVM techniques by using different features i.e. Unigrams, Bigrams, adjectives and POS. They performed three-fold cross validation and achieved 82.9% accuracy in case of SVM with Unigrams.

Govindarjan et. al. [10] proposed a new hybrid classifier based on coupling classification methods for sentiment analysis of restaurant reviews. This classifier was designed by combining NB, GA (Genetic Algorithm) and SVM using majority voting. He used 10-fold cross validation for all experiments. His results showed that their hybrid model gives better classification accuracy 92.44% than the single classifiers and GA perform better than NB and SVM GA achieve 85.30% accuracy.

Daud et. al. [4] proposed a Roman Urdu opinion mining system which took opinions in Roman Urdu as input translated them into English and then find their sentiment by matching the adjectives of opinions with manually designed dictionary and gave us the rating of the product as the output. Accuracy of this system was not so high because only adjectives cannot determine the sentiment of an opinion correctly.

Bilal et. al. [11] worked on sentiment classification of comments written in English and Roman Urdu. Their training dataset contained 300 comments (50% positive, 50% negative). They obtained these comments from a blog by using easy web extractor. They used WEKA to train three machine learning algorithms NB, KNN (K-Nearest Neighbors) and Decision Tree on this training dataset. Then, they gave testing data to these three trained models and concluded that the performance of NB in terms of Accuracy, Precision, Recall and F-measure is better than the performance of KNN and Decision Tree. In NB they achieve 97.50% accuracy, 0.974 precision, 0.973 recall and 0.975 F-measure on testing data set.

Mukhtar et. al. [12] worked on Urdu sentiment analysis. They prepared their dataset by extracting Urdu data from 151 Urdu blogs. Two human annotators annotated the data into positive, negative and neutral classes. After annotating the data, stop words were removed for dimensionality reduction. Then, they applied three machine learning algorithms SVM, KNN and Decision tree on the data by using 10-fold cross validation in WEKA. Results of these experiments are less than 50% which are not reasonable. For improvement of results, they identified important features and used only these 39 important features for representation of their data and performed experiments again. This time, they achieved 67.01% accuracy in case of KNN.

3. THE PROPOSED METHODOLOGY

Fig. 1 summarizes the architecture view of our Roman Urdu sentiment analysis system. First of all, various websites with Roman Urdu contents are crawled and saved in an opinion database. Then, the content of the dataset is reduced by applying some preprocessing. After this, important features that are usually used to apply sentiment analysis are extracted out from the preprocessed data. Later, various machine learning techniques are used on the feature sets to predict the mood of the Roman Urdu text. Finally, the predicted result is compared with the manually tagged result to evaluate the accuracy of the proposed system. The details of each step is shown in Fig. 1.

3.1 The Dataset

As mentioned earlier, there is no publically available dataset available for such purposes. Therefore, a dataset is prepared by collecting comments/opinions in Roman Urdu from different websites i.e. hamariweb, youtube, drama on line, ytpak, facebook etc. Dataset contains 806 comments of which 400 are positive and 406 are negative. We manually read all comments/opinions and assign them sentiment either positive or negative. If a comment/opinion contains more than 20% English that comment/opinion does not include in the dataset because in this study, we deal with comments/ opinions in Roman Urdu. Neutral or subjective comments/ opinions which do not have any sentiment did not include in our dataset. Dataset looks like:

{

Text: comment/opinion text, Sentiment: (pos/neg),

URL: (URL from where that comment has been taken), Domain: (Drama/Movie/Telefilm/Product/Service),

}

We have used 806 comments for training and testing purpose. The description of the dataset is summarized in Table 1.

3.2 Preprocessing

First, we manually annotate the data as negative, positive and neutral. We do this step before any processing,

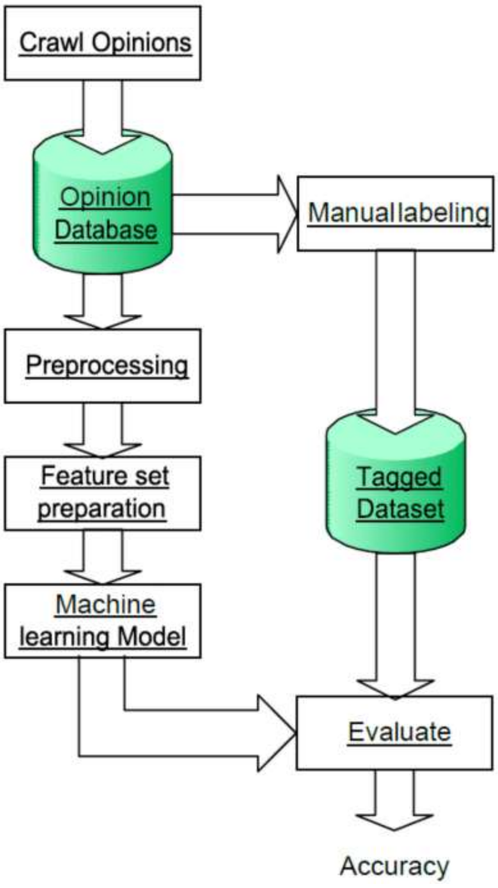

FIG. 1. THE PROPOSED ARCHITECTURE

because annotator may have difficulty in predicting the sentiment of comment/opinion due to incompleteness of sentence. When dealing with data in text, all words are not of the same importance. Secondly, they are large in numbers. Preprocessing is an important step that keeps the important words and strip off the unnecessary words. In preprocessing step, we remove unnecessary words like punctuation marks, numerical characters and stop words. The words that are common, high frequency and does not gives meaning in predicting the sentiment are called stop words. Luhn [13] was the first to introduce the concept of stop words. In our work, we have chosen the stop words manually. Data preprocessing also helps in reducing the computation time and dimensions of the data.

3.3 Feature Set Preparation

Even preprocessing the text is not enough to achieve better sentiment analysis results. The text after preprocessing is processed further to extract features that may improve the results. In literature, researcher have extracted many features for successful sentiment analysis. Here, we have chosen eight different feature sets, which are combined from various simple text features. First, we define the simple features, and then we will present our eight features sets. These eight feature sets are also commonly used in the sentiment analysis of text in English language [14].

- **N-Gram:** An n-gram is a contiguous sequence of words from a given text. when value of n is 1,it refers to Unigram, when value of n is 2, it refers to Bigrams and so on. For example, I live in Pakistan, Unigrams are "I", "live", "in", "Pakistan" and Bigrams are "I live", "live in", "in Pakistan" etc.

- **TF-IDF:** A statistical measure to determine the importance of words in a documents. It consists of two terms, TF and IDF, which are further defined as:

$$-\mathrm{TF}(t) = \ln\left(\frac{\text{Total Number t Appears in a Documents}}{\text{Number of Terms in the Documents}}\right) \quad (1)$$

$$-\mathrm{IDF}(t) = \ln\left(\frac{\text{Total Number of Documents}}{\text{Number of Documents with Term t in it}}\right) \quad (2)$$

- **OneR Attribute:** It use simple association rules, to find out only one main attribute involved in the prediction

- **Principal Component:** An axes which gives the data the maximum variation is called Principal Component.

- **Gain Ratio Attribute:** It is the ratio of information gain to intrinsic information. The purpose of using this ratio is to reduce bias toward multi-valued attributes.

TABLE 1. THE DATASET DESCRIPTION

Domain	Comments	Positive	Negative
Drama	690	356	344
Movie	43	23	20
Telefilm	23	11	12
About any topic	50	10	40

The eight combined feature sets used in our experiments are as follows:

(1) Unigram

(2) Unigram +TF-IDF

(3) Unigram + TF-IDF +OneRAttribute

(4) Bigram

(5) Bigram +TF-IDF

(6) Unigram + Bigram +TF-IDF

(7) Unigram+Bigram+TF-IDF+Principal Component

(8) Unigram+Bigram+TF-IDF+Gain Ratio Attribute

4. EXPERIMENTS

We have trained three machine learning algorithms on our dataset namely NB, LRSGD and SVM for binary classification (predict sentiment of comments/opinions either positive or negative). All three algorithms have been applied on preprocessed dataset by selecting eight feature sets and obtained results, hence a total of 24 experiments are performed. We have two main objectives to perform all the experiments. First, to select the best feature set that will give the best sentiment analysis results. Second, to find out the best machine learning algorithm for this purpose. The experiments are performed in data mining tool named WEKA. To achieve consistent results, 10-fold cross-validation is used in all the experiments. As mentioned earlier, that dataset is balanced, therefore, accuracy as an evaluation metric will suffice. It is defined as the fraction of prediction our model got right.

$$\text{Accuracy} = \left(\frac{\text{Correct Predictions}}{\text{Overall Predictions}}\right) \tag{3}$$

4.1 Correct Predictions

Results and Discussion: The experimental results of three machine learning algorithms applied on eight feature sets is shown in Table 2. The bold numbers in each row presents the maximum accuracy for the three machine learning algorithms. Table 2 shows that Unigram gives better results than Bigram for all the algorithms, even when they are augmented with other features. This is consistent with sentiment analysis of text in English too [14]. The possible reason that Bigram performs poorly than Unigram is that in Bigram, number of features explode significantly, that may hurt the machine learning algorithm to learn properly. The results are improved further when both the features are combined together. The possible reason of improved results when Unigram is combined with Bigram is that it combines the strengths of both by incorporating contextual information as well as word level information. The best results (87.22% accuracy) are obtained, when Unigram and Bigram are combined together along with TF-IDF (Unigrams + Bigrams + TF-IDF). When Gain Ratio Attribute is further added (Unigram + Bigram+ TF-IDF + Gain Ratio Attribute), it does not add to benefit. Whereas, when Principal Component is added (Unigram + Bigram + TF-IDF +Principal Component), it reduces the accuracy. Hence, the best feature set that produces the best results is Unigram+ Bigram + TF-IDF. The best machine learning algorithm is SVM, which outclass LRSGD and NB in 6 out of 8 feature sets. The best result (87.22% accuracy) is also due to SVM. The accuracy of LRSGD is always better than NB for all the eight feature set. It can be concluded that the best feature set is Unigram+ Bigram + TF-IDF and best algorithm to perform sentiment analysis in Roman Urdu is SVM.

TABLE 2. RESUTS

Features	NB (%)	LRSGD (%)	SVM (%)
Unigram	79.40	83.87	83.62
Unigram + TF-IDF	77.91	85.85	85.98
Unigram + TF-IDF + OneR Attribute	77.91	85.85	86.22
Bigram	72.82	73.82	75.55
Bigram + TF-IDF	72.45	73.94	74.93
Unigram + Bigram + TF-IDF	78.41	86.72	87.22
Unigram + Bigram + TF-IDF +Principal Component	65.01	81.01	77.04

5. CONCLUSION

In this paper, we perform sentiment analysis of Roman Urdu comments/opinions by using most popular machine learning techniques NB, LRSGD and SVM. Since, no publicly available Roman Urdu opinion/comments dataset was present, we prepared our own dataset after scraping from various websites. Many experiments have been performed by using various features very carefully in order to improve the performance of all these three classifiers. Accuracy performance measure is used in order to evaluate classifiers. The experimental results show that LRSGD and SVM perform much better than NB. There is a very little difference in performance of both LRSGD and SVM but SVM gives best results on our dataset when we used {Unigram+Bigram+TF-IDF} features set and achieve 87.22% accuracy.

6. FUTURE WORK

In future, we will extend our dataset further and will cover more domains. Furthermore, we are also planning to use deep learning approaches to solve this problem.

ACKNOWLEDGEMENT

Authors are grateful to Prof. Dr. Syed Mansoor Sarwar, Principal, Punjab University College of Information Technology, Lahore, Pakistan, for providing infrastructure to perform extensive experiments.

REFERENCES

[1] Bhonde, R., Ingulkar, B.H., and Pande, A., “Sentiment Analysis Based on Dictionary Approach”, International Journal of Emerging Engineering Research and Technology, Volume 3, pp. 51-55, 2015.

[2] Chen, P., Wu, S., and Yoon, J., “The Impact of Online Recommendations and Consumer Feedback on Sales”, Proceedings of International Conference on Information Systems, Association for Information Systems AIS Electronic Library, pp. 711-724, December, 2004.

[3] Chevalier, J., and Mayzlin, D., “The Effect of Word of Mouth on Sales: Online Book Remarks”, Indian Journal of Computer Science and Engineering, Volume 43, pp. 345-354, 2006.

[4] Daud, M., Khan, R., and Daud, A., “Roman Urdu Opinion Mining System (Ruomis)”, Computer Science & Engineering: An International Journal, pp. 1-9, 2014.

[5] “Hindustani,” https://www.encyclopedia.com/literature-and-arts/language-linguistics-and-literary-terms/language-and-linguistics/hindustani, July, 2017.

[6] Tripathy, A., Agrawal, A., and Santanu, R., “Classification of Sentimental Reviews Using Machine Learning Techniques”, Proceedings of 3rd International Conference on Recent Trends in Computing, pp. 821-829, Elsevier B.V., 2015.

[7] Shaziya, H., Kavitha, G., and Zaheer, R., “Text Categorization of Movie Reviews for Sentiment Analysis”, International Journal of Innovative Research in Science, Engineering and Technology, Volume 4, pp. 11255-11262, 2015.

[8] Kalaivani, P., and Shunmuganathan, D., "Sentiment Classification of Movie Reviews by Supervised Machine Learning Approaches", Indian Journal of Computer Science and Engineering, Volume 4, 2013.

[9] Pang, B., Lee, L., and Vaithyanathan, S., "Thumbs Up? Sentiment Classification Using Machine Learning Techniques", Proceedings of ACM-ACL Conference on Empirical Methods in Natural Language Processing, 2002.

[10] Govindarajan, M., "Sentiment Analysis of Restaurant Reviews Using Hybrid Classification Method", International Journal of Soft Computing and Artificial Intelligence, Volume 2, pp. 330-344, 2014.

[11] Bilal, M., Israr, H., Shahid, M., and Khan, A., "Sentiment Classification of Roman-Urdu Opinions Using Naive Bayesian, Decision Tree and KNN Classification Techniques", Journal of King Saud University-Computer and Information Sciences, Volume 28, pp. 330-344, 2016.

[12] Mukhtar, N., and Khan, M., "Urdu Sentiment Analysis Using Supervised Machine Learning Approach", International Journal of Pattern Recognition and Artificial Intelligence, Volume 32, pp. 1851001-1851015, 2017.

[13] Luhn, H., "The Automatic Creation of Literature Abstracts", IBM Journal of Research and Development, pp. 159-165, 1958.

[14] Liu, B., "Sentiment Analysis and Opinion Mining", Morgan & Claypool Publishers, 2012.