Sentiment Analysis of Urdu Language on different Social Media Platforms using Word2vec and LSTM

Sajadul Hassan Kumhar
Research Scholar Department of Computer Science, Sri Satya Sai University of Technology & Medical Sciences, Sehore (MP), India
Email: drsajadulhassan@gmail.com
Mudasir M Kirmani
Assistant Professor Computer Science, SKUAST-Kashmir, Srinagar (JK), India.
Email: mmkirmani@skuastkashmir.ac.in
Jitendra Sheetlani
Professor Computer Science, Sri Satya Sai University of Technology & Medical Sciences, Sehore (MP), India
Email: dr.jsheetlani@gmail.com
Mudasir Hassan
Research Scholar (Geography), Centre of Central Asian Studies, University of Kashmir, Srinagar (JK), India
Email: drmudasirhassan@gmail.com

Abstract

Sentiment analysis is the process to analyze the opinions, emotions or sentiments regarding a review, comment, organization, firms or some event etc. with the introduction of social media people tend to share their sentiments, opinions, emotions and ideas through them. Urdu which is a dominant language in Indian sub-content. Most of people tend to share their sentiments using Urdu as one of main language on these social media sites. In this paper the Urdu text available on different social media platformswill be distributed into their vector forms by using the Word2vec model and Long Short-Term Memory Units will be utilized for text classification and SoftMaxfunction will be used as an activation function in LSTM. This SoftMax function has been used for creating sentence polarity of positive, negative or neutral attribute. In this research work the whole process has been used with recurrent Neural network.

Keyword: Sentiment analysis, sentiments, opinions, emotions, Urdu, Social Media, Word2vec, Long Short-Term Memory, Classification, SoftMax, Polarity, Recurrent Neural Network.

1 Introduction

Humans communicate using natural language as a means of communication while expressing or sharing of Information. With the advent of social media platforms people tend to share their ideas, feelings, emotions, opinions and sentiments through these social media sites. Preferably

people share this information in the form of written texts. Multiple or mixture of languages are used more and more these days on these social media sites. Mixed Multilingual and mixture of languages is the common trend among large nation like India. In Indian sub-content Urdu language, which is used by most of the people while expression of their ideas, feeling and emotions on these social media sites.Due to proliferation of these social media sites the Industry Surroundings for sentiment analysis also flourished. Sentiment analysis is used for analyzing people's opinion, attitudes and emotions regarding an entity, service, organization. Sentiment analysis is sometimes referred as opinion mining, emotion mining or affective analysis etc. however in comprehensive way all terms are referred as sentiment analysis or opinion mining. As long as industry is concerned sentiment analysis is mostly frequently being used. The academics uses the sentiment analysis and opinion mining interchangeably. They both share one and the same domain of cramming work. The sentiment analysis term first time used by T. Nasukawa and Y. Jeonghee [16] (2003). The term opinion mining has shown its appearance in the research work of K. Dave et al [7] (2003). although, thefull-fledged research on the sentiment and opinion analysis appeared in the research work of S. Das and M. Chen [12,13] (2001), B. Pang et al [4] (2002), S. Morinaga et al [14] (2002), R. M. Tong [11] (2001), P. D. Turney [10] (2002) and J. Wiebe et al [6] (2000). Although Natural language processing and linguistics has prolonged records.However, Undersized research work has been carried regarding sentimentsandpeople's opinion prior to 2000. Following 2000 the research domain got momentum and became very active research domain area. Since, the information on social media sites is quite large and processing the relevant information for generation of word embeddings to this mixed social media text is a tough, tedious and cumbersome process. However, one solution to the problem is that the mixed language will be converted into a monolingual text whereby the word embedding generation will be carried outfor representing the word in vector form and perform sentiment analysis on it.This provides strong motivation for research for sentiment analysis on text present on these social media sites. However, it poses certain challenges in the research problem which had never been studied before. The research paper will discuss and define these complications and also defined the current and more advanced techniques to solve the process of sentiment analysis. As you are aware in the record of human history a huge amount of opinionated data is added to these social media platformson each passing day and is available for analysis. Without this text dataavailable on these social media platforms, alot of research would have been impossible. It is not surprise, the beginning and the prompting growth of sentiment analysis accord and co-occur with that of social media. As a matter of fact, the opining mining and sentiment analysis is in the heart of social media exploration investigation and research. In the research paper word embeddings generations will be carried out on Urdu monolingual words into dense vector representation with lower dimensionality by using CBOW model of word2vec techniques and later the LSTM of recurrent neural network will be used for classification. At last SoftMax function will be used for generating the sentence polarity of negative, positive or neutral which will help for sentiment analysis on Urdu sentences.

2 Related Research

In the literature great deal of sentiment analysis efforts have been attempted, however there are short research studies which directly supports to the study field. Therefore, in this thesis only prominent, specific and relevant studies are presented. The research studies, papers, seminars, conferences which one or other way support to study field follows, S.-M. Kim and E. Hovy (2006) [15] proposed a system for sentiment analysis using the maximum-entropy model to train the results to subsequently extracts opinions and developed a mapping of subjective lexicon to other languages by using machine translation system and subjective analysis system for English to other languages. The model proposed don't provide better results for sentiment analysis on debates about political and social agendas for opinion mining and is also not able to predict better results on unannotated data set. A. Balahur and M. Turchi(2012) [1] proposed a machine learning model using support vector machine for sentiment analysis on French, Germen and Spanish languages using different engines for machine translation that including Bing, Moses and Google. The proposed research model has not showed better performance for translation. The translation is either noisy or shows increased variance on the dataset. The classifier of the proposed model hasn't learntthe information correctly for positive and negative classes.D. Vilares et al. (2015) [5] proposeda polarity classification and machine learning model. In the proposed model. The classification was performed on twitter dataset on different languages by using machine learning techniques. In the proposed model three machine learning techniques were introduced which include monolingual model for opinion mining, monolingual pipeline for detection of languages and multilingual model to join the two monolingual models. The research model proposed hasnot performancewellbecause of minute presence of Spanish words in the available corpus. It has been alsonoticed by annotators that Spanish terms confer a larger reoccurrence of grammatical errors relatively the English ones. A. Tripathy et al(2015) [3]proposed a model for sentiment analysis on movie reviews in English language using Machine Learning Techniques, furthermore the pre-processing of data has been carried out by separating top words, characters at punctuation and numerical characters. Inaddition, in the proposed model, a numerical matrix, Term Frequency-Inverse Document Frequency (TF-IDF) were generated by making use oflabelled polarity dataset obtained from reviews of movies, in which rows constitutes reviews and columns act for features by using Machine learning algorithms (NB, SVM) for training the model. The proposed research model has weaker results for Nave Bayes and for TF-IDF on sentiment analysis than support Vector Machine. M. Abdalla and G. Hirst (2017) [9] proposed matrix translation model to infer and predict cross-lingual sentiments. In the proposed the computation done on matrix in order to transform vector space of one language into the vector space of another language. In addition, the proposed model observed that sentiment is maintainedexactly even sub-para translation also maintains fine-grained sentiment information between languages. The research model proposed hasn't improved the accuracy of sentiment regression and isn't able to fine-grain of cross lingual document in sentiment of words in over time of single language and prediction of next word is too poor. L.-C. Yu et al. (2017) [8] proposed a vector refinement model for words. The word

vector refinement model is used to clarify and refine the existing pre-trained word vector using real valued sentiment analysis intensity score maintained by sentiment lexicon. Furthermore, the proposed model is applied for pre-training word vector refinementby utilizing the Word2vec and GloVe models. The proposed model has not evaluated on the method other than SST dataset. C. A. Panday et al. (2017) [] proposed metaheuristic and optimal cluster-head model. The model is works on Cuckoo search and K-means model. The model is practiced and adapted for sentiment analysis on twitter dataset. The research model proposed hasn't improved accuracy furthermore the model hasn't improved the sarcasm and irony tweets.Z. Jianqiang and G. Xiaolin (2017) [17] proposed a model which works on deep convolution neural network. The proposed model is used for sentiment analysis on twitter social media tweets by using unsupervised learning and word embeddings to global vector representation of words. These vector word representations are used with n-gram with prior polarity score of features to shape a set of sentiment feature on tweets from twitter dataset. The sentiment features sets are integrated with CNN (Convolution Neural Network) for training and deep learning to predict labels of sentiment classification however the research model has not achieved much accuracy.Z. Sharf and S. U. Rahman (2018) [18] proposed a model for *Security* Neural Network Based sentiment analysis for large set of corpuses in Roman-Urdu from social media sites. The corpus is cleaned, lexically normalized for standard representation of words and performs part of speech tagging for better identification. The research model proposed produces low values on standardization of words and are consequently missed by the tagger. A. Rafique (2019) [2] proposed a model forsentiment and opinion analysis. In the proposed model the sentiment analysis is done on comments and opinions in Roman Urdu dataset by using three supervised learning algorithms. The three supervise algorithms used in the proposed model are Logistic Regression with Stochastic Gradient Decent (LRSGD), Naïve Bayes (NB) and Support Vector Machine (SVM). The research model proposed has not extended the dataset to more domains and no deep learning approach have been used for sentiment analysis.

3 Sentiment analysis on Urdu Words

As mentioned before, sentiment analysis is the exercise of obtaining the polarity of sentences and taking the words as the sequence to the input. These sequence of word which are input decides whether the sentence is negative, positive or neutral. In Research tasks are divided into five components viz. training of word vector generation model, creating of ID matrix for training set, RNN along with LSTM units for creation of graphs andsentence classification for training and testing.

3.1 Loading Data

First of all, word vectors will be created to Urdu Words using the pre-trained vector model. In the research study the pre-trained word vector model will be developed by using the Skip-Gram of Word2vec model. The developed model will contain 40,000 Urdu word matrix with a dimensionality of 50. In the research study two different data structure will be imported. The

imported data structure will contain python list having 40,000 words and another embedding generated word vectors of 40,000 x 50-dimensional embedding matrix. The embedding matrix will hold all the values of word vectors. In order to make it sure that everything is in list, the research study model examines the embedding matrix and dimensions (Size) of vocabulary list. The research study model can search the words and access its corresponding vector through embedding matrix.

Now the research study has vectors, In the research studythe first task is to take sentence as input and then create its vector representation. E.g. we have a sentence

Mai	nay	sochaa	film	ahchiooardilchseaptehe			
میں	نے	سوچا	فلم	اچھی	اور	دلچسپ	تھی

In order to obtain vectors to words, the embedding look-up function has been used to take two arguments one for embedding matrix and another one is ID's to one and all of the words. These ID's can be introspection and viewedas the integer representation of training set. The data pipeline can be illustrated in figure 1.1

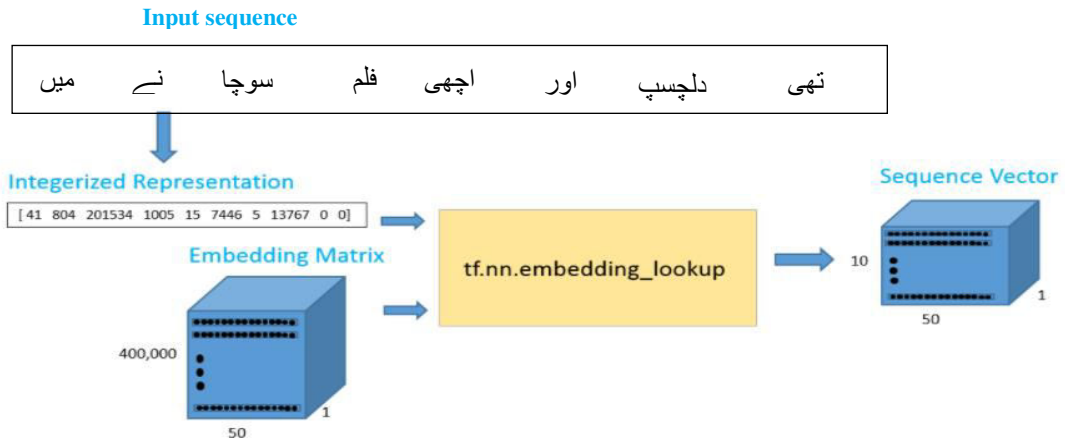


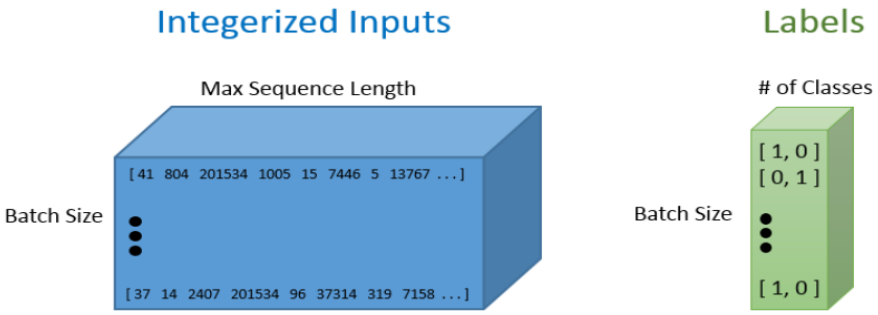
Figure 1.1: showing the Input sequence of words and producing sequence vectors

In the Figure 1. The pipeline produces output sequence vector of dimensionality 10×50. The output sequence of 10×50 must contain50-dimensionaldistribution representation of word vectors for each 10 words in that sequence. The research study model must visualize the type of data before creating the ID matrix for whole training dataset. This processing evidently will help in determining the best values by setting maximum sequence length. The training dataset that is going to be used in the research study is the social media review dataset.

3.2 RNN for classifying the review data

Now the research study model is ready to classify the dataand for creating the graph. For classifying and creating the graph some important hyper-parameters are required which includessize for each batch, number of Long Short-Term Memory units, number of classes

output by the systemand the number of times iterations required for training. In order for classifying the graph the research study requires two placeholders, one is used for input into the network and other is used for creating labels. The importance of the placeholder lies in that they are used for understanding the dimensions. The label placeholder constitutesa set of values which could be either [0,1] or [1,0] depending on one and all training example to be either positive or negative. Also,the integerized representation of training data of training example represents each row is the integerized input placeholder that is include in thebatch as shown in figure 1.2.



Now we have the data in the format required to have, and in the research study this input is fed into the LSTM network. For this purpose, the research model will use a function that takes an integer which is required for LSTM units. The aforementioned is one of the hyper-parameters that takes a part for tuning to figure out one of the admirablevalues. The research model then wraps that Long Short-Term Memorycell in a dropout layer to head off and prevent the network from overfitting.

3.3 SoftMax

SoftMax is a functionalso named assoftargmax. The SoftMax is a normalized exponential function which is used to calculate events of distribution of chances over 'n' events. The SoftMax function generalizes the logistic function to multiple dimensions. The function calculates the chances of every target class over all possible classes. later the calculated chance of probabilities is important for determining the target class for given inputs. One of the biggest advantages of using the SoftMax function is that it produces the output probability that ranges the values in-between 0 to 1. The function extractsan input of Z vector of K real numbers. The functionstandardizes it to a probability distribution which consists of K probabilities. The K probability is proportional to the exponents of the number which is input to the function. Before pertaining the SoftMax function, the output produced by vector components a range of values which are either negative, or will be greater than one, or probably may not sum to 1. However, incorporating and utilizing the services of SoftMax function the components adds up to 1 and the values calculates in the interval of (0,1). In that way the function interpreted them as probabilities. A SoftMax allows the neural network to run multiclass function. SoftMax layer is good to determine multiclass probabilities. However, there are some issue with this as the classification process become tedious when the number of classes grows. In that scenario candidate sampling can be more effective around the work.

4 Results and Discussion

sentiment analysis which was carried out by using the LSTM recurrent neural network model. In the LSTM the SoftMax was used as activation function for classifying data into neutral, Positive and negative opinions. The model showed motivational results for our Urdu dataset. In furtherance of evaluationand access the classification performance for which F-Measure, Precision, Recall and accuracy evaluation metrics were used. These evaluation metrics are utilized to differentiate different classifiers such as NB (Naïve Bayes), ELM (Extreme Learning Method) and LSTM (Long Short-Term Memory). It is pertinent to mention that word2vec is applied for all classifiers. For our Urdu dataset, the measure of performance on three different classifiers is shown in fig. 1.3.

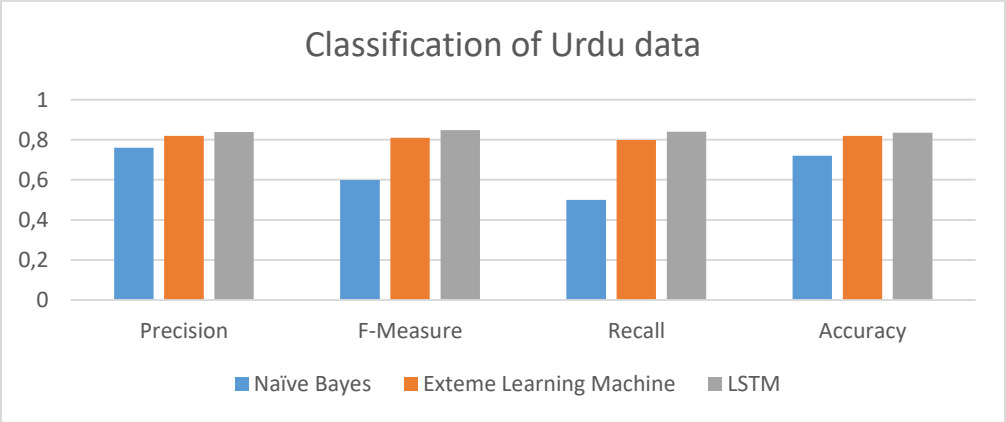


Fig. 1.3: Showing the performance measure by three classifiers for Urdu data.

On careful investigation on the performance measure shown in the fig. 5.8 above, LSTM achieved the best performance among all the three classifiers with a F-Measure of 0.849. Naïve Bayes shows worst performance because of high-rise positive false rate which is clear indication for better performance on neural network particularly for deep neural networks.

5 Conclusion

Sentiment inspection oranalysis is utilized to examine people's opinion, reviews, evaluations, sentiment toward entities, individuals, organizations and topics etc. Due to emergence of social media people tend to express their sentiments through them. In that scenario opinion mining became a necessity to see what are the sentiment that people expressed or stated to say on social media sites. In the research paper monolingual Urdu dataset was produced by translation English into Urdu and transliterating Roman Urdu into Urdu. In the research paper the skip gram model of Word2vec has been for embedding generation to Urdu words and the SoftMax function of LSTM model has been used to classify and generate the sentence polarity for sentiment analysis.

6 Futuristic Scope

In future the Model can be directly applied to any kind of text without translation and transliteration to form monolingual textand perform sentiment analysis on that corpus.

Acknowledgement

The author would like to thank to Social media scientists, Urdu language experts and experts from computer science, Dr. JitendraSheetlani, Dr. Mudasir M Kirmaniand Dr. Riyaz Ahmad Kumar for valuable support and guidance.

References

[1] A. Balahur, and M. Turchi, "Multilingual Sentiment Analysis using Machine Translation", In *Proceedings of the 3rd Workshop Association for Computational Linguistics* pages 52–60, Jeju, Republic of Korea, 12 July, 2012.

[2] A. Rafique, M. K. Malik, Z. Nawaz, F. Bukhari, and A. H.Jalbani, "Sentiment Analysis for Roman Urdu",*Mehran University Research Journal of Engineering & Technology*, Vol. 38, No. 2, Page 463-470, April 2019.

[3] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of Sentimental Reviews Using Machine Learning Techniques", In *Proceedings of 3 International Conference on Recent Trends in Computing*, pp. 821 - 829, Elsevier B.V. 2015

[4] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques", In *Proceedings of Conference on Empirical Methods in Natural Language Processing* (EMNLP-2002). 2002.

[5] D. Vilares, M. A. Alonso, and C. Gomez-Rodriguez, "Sentiment Analysis on Monolingual, Multilingual and Code-Switching Twitter Corpora", In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA*), pages 2–8,Lisboa, Portugal, 17 September, 2015.

[6] J. Wiebe,"Learning subjective adjectives from corpora", In *Proceedings of National Conf. on Artificial Intelligence* (AAAI-2000). 2000.

[7] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews" In *Proceedings of International Conference on World Wide Web (WWW2003)*. 2003.

[8] L.-C. Yu, J., Wang, K. R. Lai, and X. Zhang, "Refining Word Embeddings for Sentiment Analysis",In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics,* pages 534–539 Copenhagen, Denmark, September 7–11, 2017.

[9] M. Abdalla, and G. Hirst, "Cross-Lingual Sentiment Analysis Without (Good) Translation", In *Proceedings of the 8th International Joint Conference on Natural Language Processing*, pages 506–515, Taipei, Taiwan, November 27 – December 1.

[10] P. D. Turney, "Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews", In *Proceedings of Annual Meeting of the Association for Computational Linguistics (*ACL-2002). 2002.

[11] R. M. Tong, "An operational system for detecting and tracking opinions in on-line discussion", In *Proceedings of SIGIR Workshop on Operational Text Classification*. 2001.

[12] S. Das, and M. Chen, "Yahoo! for Amazon: Extracting market sentiment from stock message boards", In *Proceedings of APFA-2001*. 2001.

[13] S. Das, and M. Chen, "Yahoo! for Amazon: Sentiment extraction from small talk on the web", In *Management Science*, 2007. 53(9): p. 1375-1388.

[14] S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima, "Mining product reputations on the web", In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD-2002). 2002.

[15] S.-M. Kim, and E. Hovy, "Automatic identification of pro and con reasons in online reviews", In *Proceedings of the COLING/ACL Main Conference Poster Sessions*, pages 483-490. 2006

[16] T. Nasukawa, and Y. Jeonghee, "Sentiment analysis: Capturing favorability using natural language processing", In *Proceedings of the KCAP-03, 2nd Intl. Conf. on Knowledge Capture*. 2003.

[17] Z. Jianqiang, and G. Xiaolin, "Deep Convolution Neural Networks for Twitter Sentiment Analysis", In *IEEE,* 2017.

[18] Z. Sharf, and D. S. U. Rahman, "Performing Natural Language Processing on Roman Urdu Datasets", In *IJCSNS International Journal of Computer Science and Network Security* VOL.18 No.1, January, 2018