SUPPLEMENT ISSUE

IIOAB JOURNAL
ISSN: 0976-3104

ARTICLE

ANALYZING SENTIMENT IN INDIAN LANGUAGES MICRO TEXT USING RECURRENT NEURAL NETWORK

Shriya Seshadri*, Anand Kumar Madasamy, Soman Kotti Padannayil

Center for Computational Engineering and Networking, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, Amrita University, INDIA

ABSTRACT

This paper aims at improving the system which is submitted to the shared task on Sentiment Analysis in Indian Languages (SAIL2015) at MIKE 2015. In this work the tweets are classified into three polarity category namely positive, negative and neutral. Twitter data of three languages namely Tamil, Hindi and Bengali are already provided by SAIL 2015 task organizers as we have participated in the contest. Recurrent neural network is used for analyzing the sentiment in the tweets. The system performs well for recurrent neural network when compared with the system submitted to the shared task as the accuracy of the system had increased. This is due to the fact that the recurrent neural network concentrates more on language specific feature. In training, the recurrent neural network tries to learn based on the error that are generated as intermediate output. By this way the network seeks to pursue sentiment oriented feature which improves in analyzing the sentiments on tweets. We have obtained a state accuracy for the proposed system, where we achieved an accuracy of 88%, 72.01% and 65.16% for Tamil, Hindi and Bengali languages respectively for SAIL 2015 dataset.

KEY WORDS
Sentiment Analysis; Recurrent Neural Network; Polarity measure; Indian Languages

Published: 14 Nov 2016

***Corresponding Author**
Email:
shriyaseshadrik.r@gmail.com,m_anandkumar@cb.amrita.edu
Tel.: + +91 (422) 268 5594
Fax: + +91 (422) 268 6274

INTRODUCTION

Sentiment analysis is an improving and exciting field in language processing area, as sentiment plays a very important role in day to day life. Every individual depends on others opinion for many activities such as getting things from a shop, for watching movies etc. Such opinion or reviews are regularly commented on social media sites such as Facebook, Twitter, Google plus etc. Every individual are into social media for socializing with old friends, getting new friends and to entertain themselves. Social media are providing support for many companies such as Marketing, Sales, Advertising etc. Hence many companies are into social media in search of new customers for buying their product, knowing about the opinions of customers such that to improve the product based on the expectation of customers. The social media provide a wealthy textual data which contain many hidden information or opinion about the product to leverage for a fierce edge. For example, the marketers will riddle the big amount of social media information to find and information and interesting patterns, understand what their competitors area unit doing and also the means the trade is dynamic, and use the findings and improved understanding to attain competitive advantage against their competitors [1].Resolution makers use the textual information for improving their product and to enhance the business effectiveness based on the opinion that are extracted from the social media sites. Hence, mining the social medial information holds a very important role and it is necessary. Many companies are into this field for improving their product and to customizing it.

Mining the sentiment from social media is a challenging task as many unpretentious words are seen recurrently. This is due to fact that these opinions are not from experts and are from common people especially native speaker. Generally the opinion of a customer about a product or a movie is expressed in a chatty way. The opinion may also contains abbreviations, emoticons, idioms, with many grammatical fallacy. Emoticons are portrayal of body language in text based message [2].Emoticons plays an important role in expressing the opinions. Many kinds of emoticons are available. These are used in places where the opinions are to be expressed with minimum number of words. For example, in Twitter one should express their opinion within 140 characters. So in most of the tweets emoticons plays a vital role. For analyzing the tweets, one should know how the emoticons are used and how one emoticons are different from others. An ample amount of work has been done in past few years in the field of analyzing the sentiment.

RELATED WORKS

In recent years, the sentimental analyzing work has been turned towards social media text. Now analyzing the sentiment on social media text is an emerging field where, many companies are much interested in knowing about their product outcome. Sruthi et al., proposed a frame work for classifying the sentiment using a competitive layer neural network where the polarity of the text are classified into positive negative and neutral [3]. Ouyang, Xi, et al. suggested a method for analyzing sentiment using word2vec and convolution neural network. In this work 7 layer architecture model is applied for word2vec and convolution neural network for analyzing the sentence level sentiment [4]. Severyn et al. developed a new model in convolution neural network from an unsupervised method to a supervised method for initializing its parameter weight so as to improve the network [5]. Chintala et al. explores the unconventional approach using neural network for analyzing the sentiment of movie and sentence polarity [6]. Dos Santos et al. proffer a new deep convolution neural network for performing sentimental analysis in short text where character to sentence level information is exploited [7]. Sharma et al. proposed a method for

THE IIOAB JOURNAL

COMPUTER SCIENCE

analyzing the sentiment using back propagation artificial neural network (BPANN). In this work along with BPANN, it uses domain knowledge which are available in sentiment lexicon [8]. Kang et al. presented an improvised Naïve Bayes algorithm for classifying the sentiments that are collected from a restaurant. These data are classified into positive and negative using a senti-lexicon where the gap between the positive and negative reviews is narrow down [9].Rao, Yanghui et al. offered two opinion topic framework to examine the sentimental analysis of the readers [10]. Mittal et al. suggested an approach for analyzing the sentiment on Hindi language where the Hindi SentiWordNet plays a vital role. Discourse and negation rule are taken into account for analyzing the Hindi sentiment [11]. Balamurali et al. explained a new approach to cross lingual sentiment analysis using wordnet senses as feature. This is a supervised sentiment classification which is used for Hindi and Marathi languages [12]. Kumar et al. presented an approach for classifying sentiment in Indian languages with the support of distributional thesaurus and sentence level co-occurrences [13]. In the work of Yadev et al., they suggested a sentiment analyzing system for health news where it is classified into positive, negative and neutral. Here for a faster processing, neural network is used to train the system [14]. Pooja et al. uses the Hindi SentiWordNet for evaluating the sentiment from Hindi movie review. Synset replacement algorithm is used in finding the polarity of words which is associated with Hindi SentiWordNet[15]. Kamal et al. participated in Sentiment analysis in Indian languages (SAIL) contest for Hindi and Bengali languages. They classified the tweets based on Multinomial Naïve Bayes where an accuracy of 50.75% and 41.20% for Hindi and Bengali respectively [16]. For Hindi opinion mining system (HOMS) Jha et al., uses Naïve Bayes classifier for exploring the sentiment in Hindi movies as positive, negative and neutral. They also used POS tagging in which adjective is used for mining the opinions [17]. Sanjanasri in her work develop a computational framework for supervised Tamil document classification. She claim RKS can be effectively alternate to the kernel for a classifiers [18]. Vinithra et al. in their work they focus on mining the feeling in microblogging site, Twitter. Here, R tool is utilized for examining the factual information [19]. Sachin et al. participated in Sentiment analysis in Indian languages (SAIL) contest where the tweets are classified into three polarity using Regularized least square method [20]. In our previous work, Naïve Bayes algorithm is used for classifying the tweets into positive, negative and neutral [21]. Reshma et al. in their work, they proposed a classification method for classifying unstructured data [22]. Arunselvan et al in their work, Tamil movie reviews are classified into positive and negative using the word frequency as one of the feature. An accuracy of 65% is obtained for feature frequency count [23].

PROPOSED SYSTEM

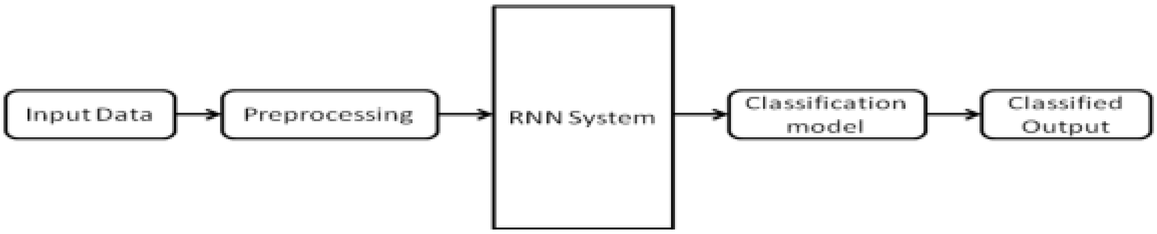


Fig.1: Sequence of the proposed system

The flow diagram of the proposed system is shown in the [Fig. 1]. As this system is an improvement of SAIL 2015, the training and testing data are given by the contest. The next step is preprocessing where the tweet id given in the input data are removed and are given to RNN system. Finally the tweets are classified into positive, negative and neutral using RNN. The system is iterated for 1000 times and the accuracy of the system is measure through F-score measure.

Input Data

This system is for improving the SAIL system. The Input of the system is already provided by the SAIL 2015 contest, as we participated in the contest. The data provided by the SAIL contest is used without any changein the size of the dataset. Both the training, development and testing data are provided earlier. SAIL is a contest mainly for Indian researchers to work on automatic sentiment analysis for improving the NLP towards Indian languages. In this contest, three main Indian languages are taken into account namely, Tamil, Hindi and Bengali. The count of training and testing data for each language is given in the [Table 1].

Preprocessing

In Data mining preprocessing plays a vital role. Due to proper preprocessing, accuracy of the system can be improved [24]. Generally, quality of the data is based on the good preprocessing [3]. The input data consist of tweet id and the tweets. In training data, the tweets are provided in three different files with the name of the respective labels. The tweet id are removed from those files with the corresponding label are provided along with the tweet separated by comma. All the files are combined to form a single file. This file is given as input to the RNN system. [Table 2] shows the tweets before and after preprocessing.

Table 1: Detailed description of Training and Testing data provided by the SAIL 2015 contest

Languages	Training Data			Test Data
	Positive	Negative	Neutral	
Tamil	387	316	400	560
Hindi	168	545	493	467
Bengali	277	354	368	500

Table 2: Tweets before and after preprocessing

S.No	Before Preprocessing	After Preprocessing
1	5,08784E+17 இந்தநிமிடத்தைமுறையாகப் பயன்படுத்தும்போதுஇன்றையநாளைமுறையாகப்பயன்படுத்திக்கொள்கிறோம். இனியகாலைவணக்கம்... http://t.co/lg8X53FJMQ	Positive, இந்தநிமிடத்தைமுறையாகப்பயன்படுத்தும்போதுஇன்றையநாளைமுறையாகப்பயன்படுத்திக்கொள்கிறோம். இனியகாலைவணக்கம்... http://t.co/lg8X53FJMQ
2	508708022852796416 सच्चाईकोअपनानाआसाननहीं...दुनियाभरसे झगड़ाकरनापड़ताहै	Negative, सच्चाईकोअपनानाआसाननहीं...दुनियाभरसेझगड़ाकरनापड़ताहै
3	508667343808258048 আমরা90 degree rocker thek।পেজটারমাধ্যমেবাংলারকসিনারিওতুলেধরাররচেষ্টাকরেছিশুরুথেকেই।সাথেবাইরের...	Neutral, আমরা90 degree rocker thek।পেজটারমাধ্যমেবাংলারকসিনারিওতুলেধরাররচেষ্টাকরেছিশুরুথেকেই।সাথেবাইরের...

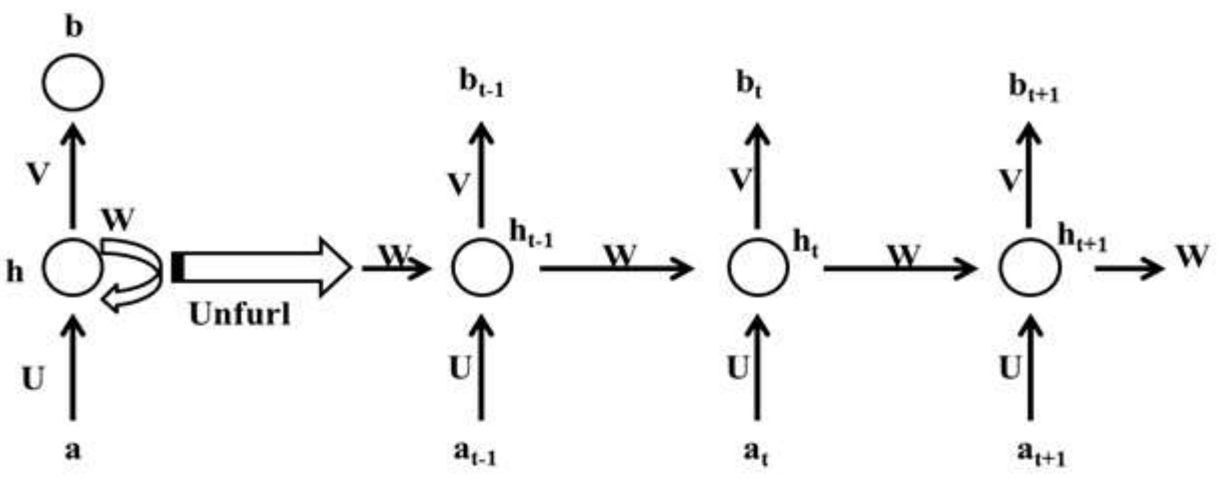
Recurrent Neural Network

Here in our work we used a simple recurrent neural network. The network consist of an input $'a'$ and an output $'b'$ with some hidden layer $'h'$. In time $'t'$ the inputs, output and the hidden layer to the network are represented as $'a(t)', 'b(t)', 'h(t)'$ respectively. Each input are in vector format. The input vector are formed by adding up the current word vector with the previous word vector. The input layer, hidden layer and the output layer are computed as follows,

$$a(t) = w(t) + h(t-1) \quad (1)$$

$$h_j(t) = f\left(\sum_i a_i(t) u_{ji}\right) \quad (2)$$

$$b_k(t) = g\left(\sum_j h_j(t) v_{kj}\right) \quad (3)$$



on
$'v'$
he
›rk
›m

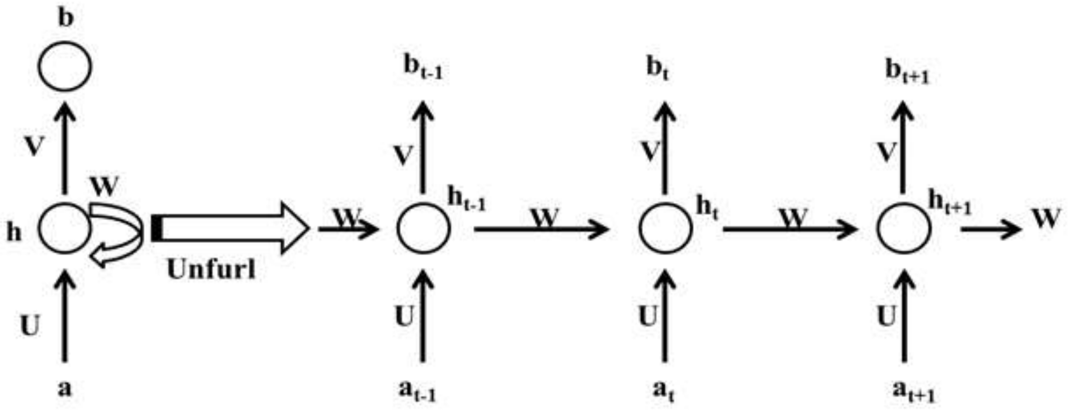


Fig. 2: Explanation of RNN network in an elaborated manner

EXPERIMENTAL ANALYSIS

The experiment is conducted on Windows-64 bit machine with 8GB RAM and i5 core processor. The SAIL (Sentiment Analysis in Indian Languages) 2015 contest is conducted as Shared task with MIKE (Mining Intelligence and Knowledge Exploration) which held at IIT Hyderabad. In the contest, twitter data of three different languages are given which is provided in the **[Table 1].** The tweets are to be classified into positive, negative and neutral. The input data contains tweet id and the corresponding tweets. As preprocessing, the tweet ids are removed, the label and the corresponding tweets are given as input to the system. The preprocessed data is given as input to the RNN system. All the three languages are preprocessed and are given to the system. The system undergoes several iterations. The accuracy of the system is obtained using F-score measure. F-score can be calculated using precision and recall. Precision is the ratio of true positive to all predicted positive and recall is the ratio of true positive to all actual positive. The accuracy of the system which are obtained using RNN is given in the below **[Table 3].[Table 4]** shows the accuracy of the system submitted to the SAIL contest with the state of art accuracy of SAIL 2015 contest with the accuracy obtained using RNN system. The system submitted to the SAIL contest is Naïve Bayes system. Extra feature are added along with the sentiwordnet for the system submitted to the SAIL. The accuracy of the system is dissipated in the second column of the [Table 4].From the table it is clear that RNN system performs well when compare with the system submitted to SAIL contest (Naïve Bayes system) and also with the state of art accuracy of the SAIL system. The accuracy of the system is improved for all the three languages and it shows a better improvement for Tamil language. [Fig. 3] represents the bar chart obtained by plotting the accuracy of the system acquired using Naïve Bayes algorithm with the state of art accuracy of the SAIL contest versus the accuracy attained using Recurrent Neural Network. From the figure-2 it is clear that the RNN system outperforms well when compared with the Naïve Bayes algorithm and we obtained a state of art accuracy for the SAIL 2015 dataset.

Table 3: Accuracy and F-Score measure of the RNN system obtained for all the three languages

S.No	Language	F-Score Measure	Accuracy
1	Tamil	0.802	88.23
2	Hindi	0.714	72.01
3	Bengali	0.644	65.16

Table 4: Accuracy of the system obtained using Naïve Bayes which is submitted to the SAIL 2015 contest and the state of art accuracy of SAIL 2015 and the accuracy obtained using Recurrent Neural Network

Languages	Accuracy of SAIL 2015(Naïve Bayes System) (%)	State of art accuracy of SAIL 2015(%)	Accuracy obtained using RNN(%)
Tamil	39.28	39.28	88.23
Hindi	55.67	55.67	72.01
Bengali	33.6	43.2	65.16

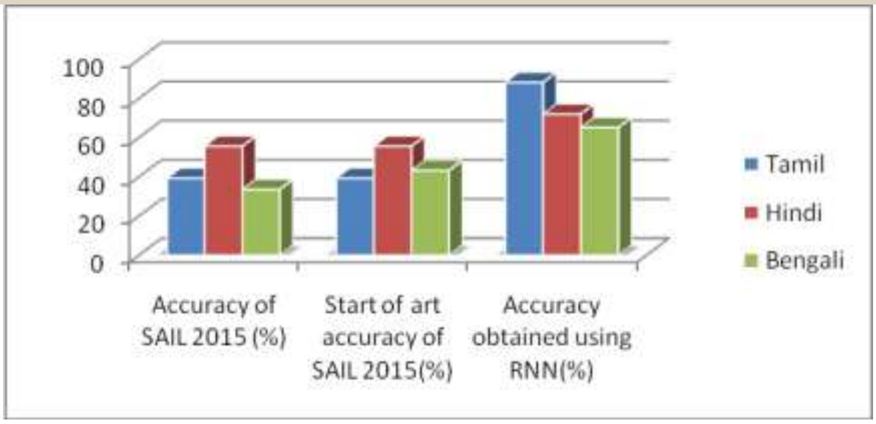


Fig. 3: Bar chart representation of accuracy obtained using SAIL 2015 system and the state of art accuracy of SAIL 2015 to the accuracy of the system obtained using RNN

CONCLUSION AND FUTURE WORK

The proposed system used to classify the tweets into positive, negative and neutral based on the content available. The system uses RNN for classifying the tweets. The accuracy of this system is higher than any other system which is submitted to the contest. An accuracy of 88.23%, 72.01% and 65.16% is obtained for Tamil, Hindi and Bengali languages respectively which is a state of art accuracy for the SAIL 2015 system. The accuracy can also be improved by using LSTM instead of RNN. The RNN has long term dependency problem but in LSTM the long term dependency problem can be overcome which leads to better accuracy. As future work, the data can be cleaned and Sentiwordnet can be added as extra feature which will leads to a better accuracy. The unsupervised data can also be added as future work.

FINANCIAL DISCLOSURE

No financial support was provided for the project

ACKNOWLEDGEMENT

I would like to express my gratitude to CEN family for their valuable support, advice and help.I avail this opportunity to express my sincere thanks to my friends, family members and staff members of my institution for their guidance, advice and encouragement at every step of this endeavor. I also express my thanks for SAIL2015 organizers.

CONFLICT OF INTERESTS

There is no conflict of interest

REFERENCES

[1] Governatori, Guido, et al. [2011] A modelling and reasoning framework for social networks policies. Enterprise Information Systems 5(1): 145-167.

[2] Ptaszynski Michal, et al. [2011] Research on emoticons: review of the field and proposal of research framework. Proceedings of 17th Association for Natural Language Processing 1159-1162.

[3] S.Suruthi et al. [2015] Neural Network Based Context Sensitive Sentiment Analysis. International Journal of Computer Applications Technology and Research Volume 4– Issue 3, 188 - 191, ISSN- 2319–8656.

[4] Ouyang Xi, et al. [2015] Sentiment Analysis Using Convolutional Neural Network. Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (CIT/IUCC/DASC/PICOM), 2015 IEEE International Conference on. IEEE.

[5] Severyn Aliaksei, et al. [2015] Twitter sentiment analysis with deep convolutional neural networks. Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM.

[6] Chintala S. [2012] Sentiment Analysis using neural architectures. New York University, New York.

[7] Santos CN, Gatti M. [2014] Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. In COLING pp. 69-78.

[8] Sharma A, Dey, S. [2012, October]. An artificial neural network based approach for sentiment analysis of opinionated text. In Proceedings of the 2012 ACM Research in Applied Computation Symposium. pp. 37-42.

[9] Kang H, Yoo SJ, Han D. [2012] Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. Expert Systems with Applications, 39(5): 6000-6010.

[10] Yanghui R, et al. [2014] Sentiment topic models for social emotion mining. Information Sciences 266: 90-100.

[11] Mittal,Namita, et al. [2013]. Sentiment analysis of hindi review based on negation and discourse relation. In proceedings of International Joint Conference on Natural Language Processing .pp. 45-50.

[12] Balamurali AR. [2012]. Cross-lingual sentiment analysis for Indian languages using linked wordnets.

[13] Ayush K, et al. [2015, December]. IIT-TUDA: System for sentiment analysis in indian languages using lexical acquisition. In International Conference on Mining Intelligence and Knowledge Exploration pp. 684-693.

[14] Yadav M, Bhojane V. Design of Sentiment Analysis System for Hindi Content.

[15] Pandey P, Govilkar S. [2015] A Framework for Sentiment Analysis in Hindi using HSWN. International Journal of Computer Applications, 119(19).

[16] Sarkar K, Chakraborty S. [2015, December] A Sentiment Analysis System for Indian Language Tweets. In International Conference on Mining Intelligence and Knowledge Exploration pp. 694-702.

[17] Vandana J, et al. [2015, July] HOMS: Hindi opinion mining system. In Recent Trends in Information Systems (ReTIS), 2015 IEEE 2nd International Conference. 366-371.
[18] Sanjanasri JP. [2015] A computational framework for Tamil document classification using Random Kitchen Sink. In 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI).
[19] Vinithra SN, et al. [2015] Simulated and Self-Sustained Classification of Twitter Data based on its Sentiment. In Indian Journal of Science and Technology, 8(24), 1.
[20] Kumar S, Sachin et al. [2015, December]. AMRITA_CEN-NLP@ SAIL2015: Sentiment Analysis in Indian Language Using Regularized Least Square Approach with Randomized Feature Learning. In International Conference on Mining Intelligence and Knowledge Exploration, 671-683.
[21] Se Shriya, et al. [2015, December]. AMRITA-CEN@ SAIL2015: Sentiment Analysis in Indian Languages. In International Conference on Mining Intelligence and Knowledge Exploration. 703-710.
[22] Reshma, U et al. [2015] Supervised methods for domain classification of tamil documents. In ARPN Journal of Engineering and Applied Sciences, 10(8):3702-3707.
[23] Arunselvan, S.J, et al. [2015]. Sentiment analysis of tamil movie reviews via feature frequency count. In International Journal of Applied Engineering Research, 10 (20), 17934-17939.
[24] Habernal I, Ptácek T, Steinberger J. [2013] Sentiment analysis in czech social media using supervised machine learning. In Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis. 65-74.
[25] Mikolov Tomas, et al. [2010] Recurrent neural network based language model. In Interspeech. 2(3).