Location Centric Review Analysis Using Social media

Stuti Gupta, K. Vimal Kumar
Jaypee Institute of Information Technology, Noida, India
{stuti.gupta, vimalkumar.k}@gmail.com

Abstract. Nowadays, social network acts as a medium for broadcasting the individual's suggestion about any product or services. The products suggestion will be uniform throughout the world, whereas, the suggestion about services depends on the location in which it was availed by the user. Review analysis provides an in-depth view about various products based on the user comments. But, in order to analyze the services in different locations, there is need for a location centric review analysis. The aim of this paper is to provide one such review analysis approach which can be helpful for many other applications such as, decision making at client level, market analysis for service based industry etc. The proposed location centric review analysis model was developed using different models for performing the review analysis and the stacked LSTM based model was found to perform better as compared with other two models. These models can be used further for developing various location centric applications.

Keywords: Sentiment analysis, Long short term memory network, review analysis, social media, location based review analysis

1 Introduction

With the adverse increase in blogging at social networks such as, Facebook, Twitter, the posting of critics or responses for various services is also increased in a faster rate. Similar to product level analysis, there is more need for service level analysis for both the client as well as the company (service provider). The business analysis can be helpful for the company to improve their service according to the feedback of the clients which can help them improve their business in that location. It also helps the user in selecting the service based on the user feedbacks. The major factor that is to be considered in this service level analysis is the location and time. There are scenario where a service being provided in a particular location (For eg., Home location of company) is much better as compared with any other locations. Similarly, the service being provided by companies gets improved over the time based on the feedback by users. In case of product review analysis, users generally provide reviews in there locally used words (transliterated in English) which may not be in the vocabulary being used for analysis. Thus, a location centric sentiment analysis will be in great demand. Due to increase in micro-blogging, much of these feedbacks/reviews are posted on social networks. Thus, there is more need for analyzing these location based service analysis using the feedbacks provided by the user on the social network. In this current research, the twitter posts are considered as the dataset of the proposed system. One of major service being used by various clients is Cab services. It has been narrowed further to Cab services based twitter post.

There are existing systems on product review analysis which uses the sentimental analysis over the reviewer comments about the product. But those product review analysis system won't suffice for the location centric review analysis. This is due to the difference in the services provided by same service provider at different location. Also, there is difference in usage of words from one location to other. The end user will be availing the services at a particular location and since there is large number of service providers, the end user has a big challenge to choose among the service providers to avail their services. The reviews about their services are being taken as a feedback by the service provider itself. But, most of the time the reviews provided by the end user does not reflect the actual feedback due to various parameters such as end user is busy with their own work. Most of the time these negative reviews are neglected by the end user while giving feedback. The service providers also provide the review analysis based on the collected feedback. But this review analysis is not transparent to the user. Thus, it gives an increase in the need for such a system which can identify the best product/service at a location.

This paper is organized and presented in the following manner: The works related to this research are summarized in the section-2. The section-3 provides complete description about the proposed system. The results of proposed system have been analyzed and detailed in the section-4. The section-5 concludes about the research and also provides a direction for the further development in this research area.

2 Related Work

A lot of research work has already been done in the field of sentiment analysis using social media data in different aspects.

In [1], authors purposed an online system for real-time collection of tweets to perform the sentiment analysis and classification. The experiment involved the use of two different classifiers, Simple Voter and Naïve Bayes. The results obtained showed that the accuracy of the system using Naïve Bayes was 81% while that of using Simple Voter was 74%.

In [2], authors collected streaming tweets based on the search term in the hashtag to identify sentiment of the user with reference to the two Indian Political parties, by user's location. Then, evaluating the tweets using Naïve Bayes classifier they achieved an accuracy of 70% and recall of 80%, while precision was 66% (as negative sentiments are harder to get right).

While [3], proposes the use of IBM's Infosphere Streams Platform (IBM,2012) for building a real time data processing infrastructure, i.e. highly scalable, enables us to write our own analysis and visualization modules to understand the dynamics of public opinion and electoral process using Twitter streaming data.

In [4], authors used three different data sets (Stanford Twitter Sentiment Corpus (STS), Health Care Reform (HCR), Obama-McCain Debate (OMD)) to do the semantic sentiment analysis of Twitter. The sentiment classification was done using unigrams, POS, Sentiment-Topic and Semantic features. According to their results obtained, Semantic classifier performed better than Sentiment-Topic classifier in case of large dataset (STS) while Sentiment-Topic classifier performed better in the topic-based datasets (HCR, OMD).

While in [5], authors have focused on sentiment analysis of tweets using binary task classifier, 3-way task classifier and a tree kernel-based model. As a result of this experiment, they have found that tree kernel and feature based model performed better than the unigram baseline model. The feature analysis revealed that the most important features were the one that combine the prior polarity of words and their parts-of-speech tags.

3 Proposed Work

Location centric review analysis has broad scope in business analysis, user level analysis etc. In business analysis, it can be used to improve their business prospects in a particular location according to the review comments of users. In case of user level analysis, the users of any service or product are in great demand for review analysis. This is due to the need for summarized analysis of huge amounts of reviews posted on social media. The social media post, specifically twitter post, generally has limitation on the maximum number of characters being used by the blogger (140 characters). Predicting the sentiment of these blogs is further a challenging thing due to the character limits. Thus, the main objective of this research is to predict the best service in a particular location based on reviews provided on social network. The user provides his location on the system and based on the location the tweets of the required service are crawled and are fed to the review analysis system. The crawling of tweets is also in an incremental manner. Before crawling, the system will keep a check on the past trained tweets and extracts only those tweets which are not used for analysis. This process is made to increase the speed of the overall system. These tweets are further fed to each of their respective review analysis models as shown in figure-1 below.

3.1 Preprocessing

The input tweets are initially cleaned to extract the text content. The cleaned text is further subjected to the process of word tokenization and spell checker. The words are segregated using the process of word tokenization and are further fed a spell checker which handles any typo errors. The cleaned text has to be encoded before feeding to the review analysis module. For this purpose, the text is label encoded to generate equivalent vectors for it. These generated vectors are further used by the review analysis module which is described in the next sub-section.

3.2 Review Analysis

The review analysis model has been designed using different recurrent neural network (RNN). Initially, model was developed using simple recurrent neural network (simple RNN). The simple RNN [6] takes sequential information but, the prediction of a network's output has dependency on its input and also on its previous input's hidden state. This is made to extract the relationship between current output and the network's previous state. This model will be more helpful in natural language processing tasks as there is dependency between one sentence and other. The RNN model is designed using sequential approach with a RNN layer and dense layer that follows the RNN layer. Since there is need for summation of all the activated neurons in the output layer, a softmax activation function is used at the output layer. But, the disadvantage of simple RNN is vanishing gradient problem. The gradient learnt at present won't be available for predicting the network output after n-time units when n is very large.

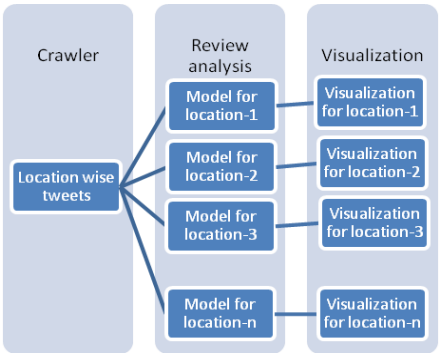

Fig. 1. Overall architecture of location centric review analysis system

To handle the vanishing gradient problem in RNN, the review analysis model is developed using a long-short term memory network (LSTM). The LSTM network [6, 7] keeps a memory to store the cell state and the hidden state. The memory is governed by four different gates – cell gate, input gate, output gate and forget gate. These gates are used to predict the need for previous cell state in the current network calculation. The LSTM model used has three layers in sequential manner and the layers used are embedding layer, LSTM layer and dense layer. Embedding layer is used to map the words in input text to a vector that will be usable for LSTM layer. The LSTM layer is followed by a dense layer which is activated using a softmax function. This LSTM network suffered from over-fitting issues. To make the model learn without over-fitting, a dropout regularization mechanism has been employed in the model. Dropout regularization mechanism [8] nullifies random percentage of neurons and allows the network to learn with other active neurons. The dropout percentage used is 40%. But, still there is scope of improving the accuracy.

In order to increase the accuracy further, a stacked LSTM network was used. In a stacked LSTM network, instead of having just one LSTM layer, there will be more than one LSTM layer in a cascaded manner. This review analysis model is designed using an embedding layer, more than one LSTM layer and a dense layer. The embedding layer is used to embed the word into a vector form. There are four LSTM layers used in this model. Each LSTM layer learns the features from the input fed to it. The first LSTM layer learns the coarse features from in the embedded output. The coarse features are passed on to the next LSTM layer, which learns features to a fine level. And the last LSTM layer learns the fine grained feature to produce the output. The output from the last LSTM layer is once again fed to a dense network. The dense network sums up all the neurons to generate the network output.

3.3 Visualization

The generated output from review analysis model is further presented according to the user requirements. In case of business/general user, the generated output is presented in a location wise manner, which can help in making any decision. Along with the location wise display, a set of keywords that influences the review analysis is also presented. To extract the keywords, the reviews are filtered based on the analysis and the top ranked keywords (excluding the stop words) in the filtered texts is displayed to the user. These keywords suggest the user about any further improvements required for future business developments.

4 Result Analysis

The proposed system has been evaluated based on three evaluation metrics – precision, recall and accuracy. Since the overall system has been implemented using various RNN models, a comparison is made between each of these models. Based on the analysis, it is found that the accuracy of stacked LSTM model is better as compared with other models. The simple RNN based model has poor accuracy due to the vanishing gradient problem. The improvement in stacked LSTM model is basically due to the increase in extraction of fine grained features as compared with the other two models.

Table 1. Results of location centric review analysis

	RNN			LSTM			STACKED_LSTM		
Location	**Accuracy (in %)**	**Precision**	**Recall**	**Accuracy (in %)**	**Precision**	**Recall**	**Accuracy (in %)**	**Precision**	**Recall**
Delhi	**65**	0.82	0.64	**73**	0.5	0.67	**91**	0.75	1
Chennai	**41**	0.8	0.36	**57**	0.33	0.5	**71**	0.67	0.67
Mumbai	**61**	0.9	0.55	**87**	0.92	0.85	**90**	0.84	0.97

The precision and recall is also improved in stacked LSTM model as compared with other two models. But, in case of single layer LSTM network, there is dip in the precision and recall for the locations Delhi and Chennai. This is due to lesser number of tweets that was available for these locations.

5 Conclusion & future work

The proposed location centric review analysis was individually developed using simple RNN, LSTM and stacked LSTM networks. It has been noticed that the performance of stacked LSTM based review analysis system is higher as compared with other two models. The precision and recall of all the models are calculated and analyzed. It is found that the precision and recall are not good for the locations in which lesser number of tweets is available on social media. In order to improve this system further, there is need for some mechanism which can handle the low resource issue. The performance of overall system can also be improved by the usage of location centric vocabulary along with their sentiments.

References

1. Al Shammari, Alaa S. "Real-time Twitter Sentiment Analysis using 3-way classifier." In *2018 21st Saudi Computer Society National Computer Conference (NCC)*, pp. 1-3. IEEE, 2018.
2. Almatrafi, Omaima, SuhemParack, and Bravim Chavan. "Application of location-based sentiment analysis using Twitter for identifying trends towards Indian general elections 2014." In *Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication*, p. 41. ACM, 2015.
3. Wang, Hao, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. "A system for real-time twitter sentiment analysis of 2012 us presidential election cycle." In *Proceedings of the ACL 2012 System Demonstrations*, pp. 115-120. Association for Computational Linguistics, 2012.
4. Saif, Hassan, Yulan He, and Harith Alani. "Semantic sentiment analysis of twitter." In *International semantic web conference*, pp. 508-524. Springer, Berlin, Heidelberg, 2012.
5. Agarwal, Apoorv, BoyiXie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. "Sentiment analysis of twitter data." In *Proceedings of the workshop on languages in social media*, pp. 30-38. Association for Computational Linguistics, 2011.
6. Sherstinsky, Alex. “Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network.” *CoRR abs/1808.03314* (2018): n. pag.

7. Hochreiter, S. & Schmidhuber, J. Long short-term memory. Neural Comput. 9, 1735–1780 (1997).
8. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. J. Machine Learning Res. 15, 1929–1958 (2014).
9. Dingqi Yang, Daqing Zhang, Zhiyong Yu, and Zhu Wang. 2013. A sentiment-enhanced personalized location recommendation system. In Proceedings of the 24th ACM Conference on Hypertext and Social Media (HT '13). ACM, New York, NY, USA, 119-128. DOI=http://dx.doi.org/10.1145/2481492.2481505
10. Omaima Almatrafi, Suhem Parack, and Bravim Chavan. 2015. Application of location-based sentiment analysis using Twitter for identifying trends towards Indian general elections 2014. In *Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication* (IMCOM '15). ACM, New York, NY, USA, Article-41, 5-pages. DOI=http://dx.doi.org/10.1145/2701126.2701129