International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 02 | Feb 2019 www.irjet.net p-ISSN: 2395-0072

Text Classification using Recurrent Neural Network in Quora

D. Yogeshwaran[1], Dr. N. Yuvaraj[2]

[2]Associate Professor, Dept. of Computer Science and Engineering, KPRIET, Tamilnadu, India
[1]Student, Dept. of Computer Science and Engineering, KPRIET, Tamilnadu, India

---

Abstract - *Text analysis plays one of the main roles in this digital era, where millions of people generates data through internet every minute. In Social network websites people shares their opinions and ideas in daily routine. These data shared by millions of people are analyzed and classified for the future prediction and better decision making. Quora is the one of the social network website where people shares their knowledge and learn. The major problem in any social network website is how to handle the toxic content in it. Quora also meets the same problem where some people misuses the purpose of the website by asking insincere questions. To tackle this problem in this paper we discussed the methodology which follows classification strategy to identify the insincere questions asked by the users. By classifying the insincere questions by users, we can avoid the misuse of the purpose of the website.*

Key Words: Text classification, Recurrent Neural Network, Toxic content analysis, Polarity detection.

1. INTRODUCTION

In the last decade growth of social networking sites has been increased tremendously. Nowadays, social networking sites where dealing with huge amount of data shared and generated by public. Millions of people express their views and opinions on a variety of topics via social network websites. Text analysis is the process of collecting the data from the resources and extracting the knowledge for the future purpose. It is important to analyses the data in social network in order to study the people's views and opinion towards the particular topic for future prediction and enhancement. It is also important to identify the toxic content in data to avoid the misuses of the particular website or social network blogs.

There are also other social network websites which are tend to meet the same problem in handle the toxic content in the data. In existing systems various models for text analysis and classification are implemented for social network websites like Facebook, Twitter and Instagram. But there is no any other models implemented for Quora, in this paper we discussed the methodology to classify the insincere questions. By classifying the insincere questions, we can avoid the toxic content in the website. Data sets are collected from the respective website and preprocessed to extract the features and the respective features are classified to encounter the insincere questions.

Quora is the well known social network where the users posts their views, opinion and seeking for answers in all topics. There are some people misuses the purpose of the website by shooting insincere questions. We have done classification on questions which are asked by people in Quora. Our proposed methodology is to detect the toxic content in the questions by extracting the features and classify the features as sincere or not. An insincere question is defined as a question intended to make a statement rather than look for helpful answers. Some characteristics that can signify that a question is insincere are non-neutral tone, is disparaging or inflammatory, isn't grounded in reality and Uses sexual content.

2. LITERATURE SURVEY

2.1 SENTIMENT ANALYSIS ON TWITTER DATA-SET USING NAIVE BAYES ALGORITHM

This paper discusses the extraction of sentiment from a famous micro blogging website, Twitter where the user posts their views and opinion. This survey does sentiment analysis on tweets which help to provide some prediction on business intelligence. This method uses Hadoop Framework for processing movie data set that is available on the twitter website in the form of reviews, feedback, and comments. Results of sentiment analysis on twitter data will be displayed as different sections presenting positive, negative and neutral sentiments.

2.2. A SURVEY OF SENTIMENT ANALYSIS TECHNIQUES

This paper presents a survey of sentiment analysis and classification algorithms. This survey concludes that sentiment classification is still an open field for research. There is a lot of scope for algorithms in it. SVM and naïve bayes are most popular algorithms for sentiment classification. Sentiment analysis of tweets is very popular. Datasets from sites like Amazon, IMDB, flipkart are widely used for sentiment analysis. Deeper analysis is required in case of social networking sites. In many cases, context consideration is very important. Therefore more research is required in this field.

2.3. A SURVEY ON TEXT CLASSIFICATION TECHNIQUES FOR SENTIMENT POLARITY DETECTION

In this survey paper, it discussed the various approaches for polarity shift detection in sentiment analysis and the overview of the sentiment analysis system. Polarity shift in the sentiment analysis can be detected by five different approaches; Dual sentiment analysis model, Polarity shift detection, elimination and ensemble approach, Term counting with polarity shifting approach, Sentence polarity shift algorithm and Sentiment classification with polarity shifting detection approach. Polarity shift detection, elimination and ensemble model can detect and eliminate all the types of polarity shifts. Hence polarity shift can be easily detected and eliminated by different polarity shift detection approaches. It helps to improve the performance of the machine learning classification algorithms.

2.4. A COMPREHENSIVE STUDY OF TEXT CLASSIFICATION ALGORITHMS.

This paper compared the different algorithms and produces some results as follows, Rocchio Classification - becomes inaccurate when the centroid of the classes does not represent its behavior well, KNN Classification - Higher time and space complexity as it stores all the instance, Noisy features degrades the classification accuracy, Naïve Bayes classification - Independence assumption of features, Rule Based Classification - Computational cost is high. Decision Tree Classification - Noise handling is bad, no online learning, over fit.

2.5 A DETAILED SURVEY AND COMPARATIVE STUDY OF SENTIMENT ANALYSIS ALGORITHMS

This study gives an idea about the processes involved in sentiment analysis. This survey concluded that hybrid approaches gives the better accuracy than the traditional approaches and algorithms. This paper compared the different papers and approaches and finally got the results as better performance in Hybrid approach. Moreover, a comparative analysis of genetic algorithm used for various tasks of sentiment analysis is also presented in the study. The results obtained from other researchers signify how the Hybrid approach is used in combination with other approaches on different datasets giving different accuracy.

2.6 TWITTER SENTIMENT CLASSIFICATION USING NAIVE BAYES BASED ON TRAINER PERCEPTION

A study done in this paper is strategy to classify tweets sentiment using Naïve Bayes techniques based on trainers' perception into three categories; positive, negative or neutral. 50 tweets of 'Malaysia' and 'Maybank' keywords were selected from Twitter for perception training. In this study, there were 27 trainers participated. Each trainer was asked to classify the sentiment of 25 tweets of each keyword. Results from the classification training was then be used as the input for Naïve Bayes training for the remaining 25 tweets. The trainers were then asked to validate the results of sentiment classification by the Naïve Bayes technique.

2.7 SENTIMENT ANALYSIS FOR THE NEWS DATA BASED ON THE SOCIAL MEDIA

In this analysis it tend to propose a replacement methodology to try and do the sentiment analysis for news data a lot of specially ,supported the social media information and social news (i.e.text and emotions text) of a happening , a Levenshtein algorithm is made to together categorical its linguistics and emotions, that lays the muse for the happening sentiment analysis. The proposed method uses Naïve Bayes and Levenshtein algorithm to determine the emotion into different categories from given social media news data.

2.8. SENTIMENT CLASSIFICATION: FEATURE SELECTION BASED APPROACHES VERSUS DEEP LEARNING

In this study, an extensive comparative study was carried out among three well-known feature selection based approaches, including word embedding features, and three popular deep learning models for document-level sentiment classification. For feature selection based approaches, selected BoW features and BoW features enriched with word embedding features were used as input into a SVM classifier. For evaluating deep learning models, CNN, LSTM, and CNN+LSTM neural network models are implemented. For three out of the four datasets, various deep learning models outperformed feature selection based approaches. However, IG+WE outperformed the all other systems in one dataset. The IG and DFS feature selection methods seemed more successful than the GI method for sentiment classification. The experimental results also indicated that better results can be obtained by initializing deep learning models with either one-hot vectors or fine tuned semantic word embeddings than the word embeddings without tuning method.

2.9 SENTIMENT ANALYSIS AND TEXT SUMMARIZATION OF ONLINE REVIEWS

This survey provides the comprehensive overview of recent and past research on sentiment analysis and text summarization and provides excellent research queries and approaches for future aspects. According to our experiment, the Naïve Bayes classification proves to be the most efficient among three algorithms for text classification of opinion mining. This work focuses only on the reviews taken from Amazon website using 3 different algorithms. The work can be extended on mining reviews from multiple website such as Flip kart, Snap deal etc. Further, to incorporate more classification algorithms to analyze their efficiency. This will help us in deciding the best text classifier in opinion mining and sentiment analysis.

2.10 MINING OPINION FROM TEXT DOCUMENTS

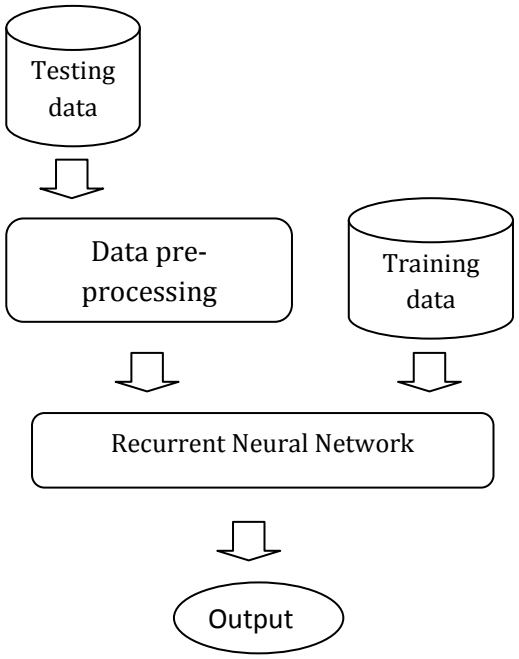
This survey presented techniques and methods that promise to enable us to get opinion oriented information from text. This research effort deals with techniques and challenges related to sentiment analysis and Opinion Mining. This paper followed systematic literature review process to conduct this survey. Their focus was mainly on machine learning techniques on the basis of their usage and importance for opinion mining. This research tried to identify most commonly used classification techniques for opinionated documents to assist future research in this area.

2.11 NAIVE BAYES CLASSIFICATION OF UNCERTAIN DATA

This document proposed a novel naive Bayes classification algorithm for uncertain data with a pdf. Their key solution is to extend the class conditional probability estimation in the Bayes model to handle pdf's. Extensive experiments on UCI datasets show that the accuracy of naive Bayes model can be improved by taking into account the uncertainty information.

3. PROPOSED METHODOLOGY

In existing systems mostly they used machine learning and data mining algorithms to classify the text. But in our proposed model we focus on Deep learning algorithm to classify the text. In many studies and researches proven that the deep learning algorithms produces better accuracy than the machine learning and traditional algorithms. In this paper we discussed the RNN algorithm to get the better accuracy. The architecture of our proposed system is shown below.



3.1 DATA COLLECTION

Data collection is a process of collecting information from all the relevant sources to find answers to the research problem, test the hypothesis and evaluate the outcomes. Data collection methods can be divided into two categories: secondary methods of data collection and primary methods of data collection.

3.2. DATA PRE-PROCESSING;

Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. There are certain steps in data pre-processing,

1. Import the libraries.

2. Import the data-set.

3. Check out the missing values.

4. See the Categorical Values. 5. Splitting the data-set into Training and Test Set.

6. Feature Extraction.

3.2.1 NLP

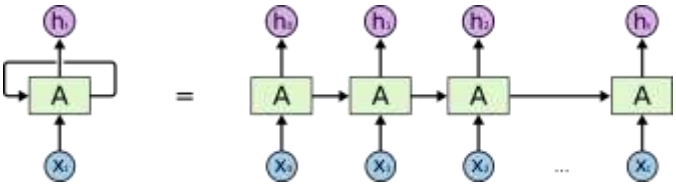
Natural language processing (NLP) is a method to translate between computer and human languages. It is a method of getting a computer to understandably read a line of text without the computer being fed some sort of clue or calculation. In other words, NLP automates the translation process between computers and humans. Traditionally, feeding statistics and models have been the method of choice for interpreting phrases. Recent advances in this area include voice recognition software, human language translation, information retrieval and artificial intelligence. There is difficulty in developing human language translation software because language is constantly changing. Natural language processing is also being developed to create human readable text and to translate between one human language and another. The ultimate goal of NLP is to build software that will analyze, understand and generate human languages naturally.

3.2.2 RECURRENT NEURAL NETWORK

Recurrent Neural Networks are one of the most common Neural Networks used in Natural Language Processing because of its promising results. The applications of RNN in language models consist of two main approaches. We can either make the model predict or guess the sentences for us and correct the error during prediction or we can train the

model on particular genre and it can produce text similar to it, which is fascinating.

The logic behind a RNN is to consider the sequence of the input. For us to predict the next word in the sentence we need to remember what word appeared in the previous time step. These neural networks are called Recurrent because this step is carried out for every input. As these neural network consider the previous word during predicting, it acts like a memory storage unit which stores it for a short period of time.



Training a typical neural network involves the following steps:

1. Input an example from a dataset.

2. The network will take that example and apply some complex computations to it using randomly initialized variables (called weights and biases).

3. A predicted result will be produced.

4. Comparing that result to the expected value will give us an error.

5. Propagating the error back through the same path will adjust the variables.

6. Steps 1–5 are repeated until we are confident to say that our variables are well-defined.

7. A predication is made by applying these variables to a new unseen input.

4. CONCLUSION AND FUTURE WORK:

Quora data in the form of opinion, feedback, reviews, remarks and complaint are treated as big data and it cannot be used directly. These data first is to convert as per requirement. In this paper we discussed the deep learning approach to encounter insincere questions in Quora. We expect deep learning algorithms will give the better performance and efficiency than the traditional approaches. This type analysis will definitely help any organization to improve their business productivity. The analysis of social network data is done on various perspectives like Positive, Negative and Neutral sentiments on data. Hence, the future scopes in this approach for the other social networking websites like Facebook, Twitter, Instagram etc

REFERENCES

1) Brindha, S., Prabha, K., & Sukumaran, S. (2016). A survey on classification techniques for text mining. 2016 3rd International Conference on Advanced Computing and Communication Systems (ICACCS).

2) Kaur, H., Mangat, V., & Nidhi. (2017). A survey of sentiment analysis techniques. 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (ISMAC).R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

3) Arunachalam, N., Sneka, S. J., & MadhuMathi, G. (2017). A survey on text classification techniques for sentiment polarity detection. 2017 Innovations in Power and Advanced Computing Technologies (iPACT).

4) Vijayan, V. K., Bindu, K. R., & Parameswaran, L. (2017). A comprehensive study of text classification algorithms. 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI).

5) Sinha, H., & Kaur, A. (2016). A detailed survey and comparative study of sentiment analysis algorithms. 2016 2nd International Conference on Communication Control and Intelligent Systems (CCIS).

6) Parveen, H., & Pandey, S. (2016). Sentiment analysis on Twitter Data-set using Naive Bayes algorithm. 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT).

7) Ibrahim, M. N. M., & Yusoff, M. Z. M. (2015). Twitter sentiment classification using Naive Bayes based on trainer perception. 2015 IEEE Conference on e-Learning, eManagement and e-Services (IC3e).

8) Ren, J., Lee, S. D., Chen, X., Kao, B., Cheng, R., & Cheung, D. (2009). Naive Bayes Classification of Uncertain Data. 2009 Ninth IEEE International Conference on Data Mining.

9) Fan, X., Li, X., Du, F., Li, X., & Wei, M. (2016). Apply word vectors for sentiment analysis of APP reviews. 2016 3rd International Conference on Systems and Informatics (ICSAI).

10) Aydin, M., & Baykal, N. (2015). Feature extraction and classification phishing websites based on URL. 2015 IEEE Conference on Communications and Network Security (CNS)

11) Stanojevic, M., & Vranes, S. (2005). A Natural Language Processing for Semantic Web Services. EUROCON 2005 - The International Conference on "Computer as a Tool."

12) Shahare, F. F. (2017). Sentiment analysis for the news data based on the social media. 2017 International

Conference on Intelligent Computing and Control Systems (ICICCS).

13) Gupta, P., Tiwari, R., & Robert, N. (2016). Sentiment analysis and text summarization of online reviews: A survey. 2016 International Conference on Communication and Signal Processing (ICCSP).

14) Khan, K., Baharudin, B. B., Khan, A., & e-Malik, F. (2009). Mining opinion from text documents: A survey. 2009 3rd IEEE International Conference on Digital Ecosystems and Technologies.

15) Li, L., Xiao, L., Wang, N., Yang, G., & Zhang, J. (2017). Text classification method based on convolution neural network. 2017 3rd IEEE International Conference on Computer and Communications (ICCC).