International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 06, June-2020

Application of Logistic Regression in Natural Language Processing

Bhartendoo Vimal
VI Sem, MCA
Department of MCA
RV College of Engineering,
Bengaluru

Dr. S. Anupama Kumar
Associate Professor
Department of MCA
RV College of Engineering, Bengaluru

***Abstract*:- Data Delicacy is one of the most important issue now a days, not only storing it requires a lot of storage space but even kills a lot of time. These generally happens in online discussion forums mostly, Quora is one of them. This paper gives an insight into handling the problem of actual duplication of questions. So, to overcome the problem Quora issued a public dataset in which users were asked to give a solution to their problem which should be time efficient and should categorize the dataset as duplicate or non-duplicate. NLP is the most important part to carry out this paper which helps in stemming of the question. Similarly, Tf-idf word vector helps in conversion of words and characters into computer understandable format. Since classifying characters by computers is not an easy task unless it's converted to binary. Then lastly applying Logistic Regression to train a model which will classify the next set of questions from itself. So, after this the forums will have cheaper data storage – storing less questions, Improved customer experience – faster responses to questions, Re-use content – if a question has been answered before it is very efficient to use the same answer for a similar question.**

Keywords:- NLP – Natural Language Processing , Pre-Processing , Model Log-Loss , Accuracy

1. INTRODUCTION

Data Duplicity is one of the important problem in text analysis which consumes more time and reduces the efficiency of the system. In any query processing system like Quora, users face certain problems like duplication of questions. Suppose, a user posts a question on the website and may be there is already a solution for the question. But there is possibility that the wordings of the question asked by user might not be similar to the one in Quora database. But even after that the user must get the answers. So in this case NLP plays an important role to find out which questions are similar and which are not. In this paper we address the problem of actual duplication of questions. In the problem understanding semantic relatedness of sentences would allow understanding of much of user-generated content on the internet, such as on Quora. In this paper we address the problem of actual duplication of questions. Solving the problem will help Quora organize and deduplicate their knowledge base.

This issue can be handled by Natural Language Processing of Text. Natural Language Processing, usually shortened as NLP, is a branch of artificial intelligence that deals with the interaction between computers and humans using the natural language. The ultimate objective of NLP is to read, decipher, understand, and make sense of the human languages in a manner that is valuable for computers.

Website like Quora, is known for providing solutions for technical questions to the users. So it has to be quite quick at responding with solutions to the question. There is a need to highlight the duplicate questions which will help in cheaper data storage – storing less questions, Improved customer experience – faster responses to questions, Re-use content – if a question has been answered before it is very efficient to use the same answer for a similar question. And after all these if it fails to provide an appropriate result then Quora loose trust so the cost of misclassification can be high.

This paper intends to give a solution to the problem of eliminating data duplicity and give a faster response to the user using Text Analytics and Machine learning. Basically, since both NLP and ML has been used so in NLP to segregate the questions into machine understandable way tokenization is done which maps each word to a unique integer index. Fuzzy is used to calculate the probability of similarity of two questions and finally tf-idf to map each word into a machine understandable format. While applying ML, Logistic Regression is best as it provides the best solution for the model while Log-loss is minimum and accuracy is high.

The paper is divided into four sections, the introduction, literature survey, proposed work and results.

2. LITERATURE SURVEY

Tim Rocktaschel, Edward Grefenstette, Karl Moritz Hermann, Tomas Kocisky, Phil Blunsom. Reasoning about entailment with neural attention. In ICLR 2016. [1] This paper proposes us a neural network which can extract and read two sentences to find out semantics of both the sentences and find out co-relation among them.

Minh-Thang Luong, Hieu Pham and Christopher D. Manning. Effective Approaches to Attention-based Machine Translation. In EMNLP,2015. [2] This paper, shows how attentional mechanism can be used to improve neural machine translation by selectively focusing on parts of source sentence. This paper examines the word source and even the subsets of the word to break it down to its root form and this technique is known as stemming.

Jonas Mueller and Aditya Thyagarajan. Siamese Recurrent Architectures for Learning Sentences Similarity. In AAAI, 2016.[3] This paper shows that the model is applied to get semantic similarity among sentences which means any sort of relevance through which one can classify the sentences. It outperforms the features that were extracted one by one.

Shuohang Wang and Jing Jiang. Learning natural language inference with LSTM. In Proceedings of NAACL, 2016.[4] In the paper the neural network perform word by word matching of the scentences. The LSTM places more emphasis on important words and even remembers important mismatches that are critical for predicting.

Gabor Angeli and Christopher D. Manning. Naturalli:Natural logic inferencefor commomscence reasoning. In NIPS,2014.[5] Reasoning is important for AI applications and that to Common-sense, in NLP or any vision and robotics tasks. A Natural Logic inference system for inferring common sense facts is proposed from a very large database of known facts. This helps a system to predict common sense facts with 91% accuracy.

Samuel R. Bowman, Gabor Angeli, Christopher D. Manning. A large annotated corpus for learning natural language inference. In EMNLP,2015.[6] Natural language processing is there for analyzing semantic representation of the dataset. How ever machine learning lacked resources to find out semantic of the data. So, SNLI is freely available which has collection of pair of scentences which help to analyze data.

Minh-Thang Luong, IIlya Sutskever, Quoc V. Le, Oriol Vinyals, Wojciech Zaremba. Addressing the Rare Word Problem in Neural Machine Translation. In ACL, 2015.[7] Neural Machine Translation (NMT) is a new approach to translate words and it has shown promising results that are much more efficient to traditional approaches. A weakness in old NMT systems is their inability to correctly translate rare words. In this paper, this paper proposes an effective technique to address this problem.

IIlya Sutskever, Quoc V. Le and Quoc V Le. Sequence to sequence learning with neural network. In NIPS, 2014.[8] Deep Neural Networks (DNNs) is one of the most effective model that can perform difficult tasks. DNN works well on large labelled training set of data. In this paper one maps the input data to a vector so that the model could learn the data effectively.

Jeffrey Pennington, Richard Socher and Christopher D. Manning. 2014. GloVe:Global Vectors for Word Representation.[9] Recent methods of learning vector representation have succeeded in finding semantic and regularity in the datasets. Vectors have been used to convert words to a computer understandable format.

Jiang Zhao, Tian Zhu and Man Lan. Enchu: One stone two birds:Ensemble of heterogenous measures for semantic relatedness and textual entailment. In SemEval, 2014.[10] This paper addresses two issues – can one solve two tasks together – feature proposed for textual entailment, is it still effective for semantic related task. For this here seven attributes, say features are extracted and common text similarity are calculated.

By the Analysis of literature survey the knowledge is gained about how various machine algorithm works and how accurate enough that algorithm would be for the project. Finding out the best suited for the project. NLP being a part of the project is used to decode the series of string into a computer understandable way.

3. PROPOSED MODEL

The following figure 1 describes proposed model. Logistic regression will be used to train the model. The data set used in this work is taken from kaggle. The dataset which has been provided by Quora has been pre-processed removing punctuations, Tokenization, removing stopwords, stemming and Vectorizing data(tf-idf). The details of the functionalities are listed below:

Tokenization – It's a process which create a large dictionary which maps each word to a unique integer index. This dictionary is then used to convert sentences from sequences of string to sequences of integers.

Stemmimg – It's a process of producing root word of a extended word phase. Stemming algorithm basically reduces the word such as “Likes, Liked, Likely, Liking” to root word “Like”.

Vectorization – It's a process of converting an algorithm from operating on a single value at a time to operating on a set of values at one time.

Tf-idf – Tf-idf is an abbreviation for Term Frequency-Inverse Document Frequency and is a very common algorithm to transform text into a meaningful representation of numbers. The technique is widely used to extract features across various NLP applications.

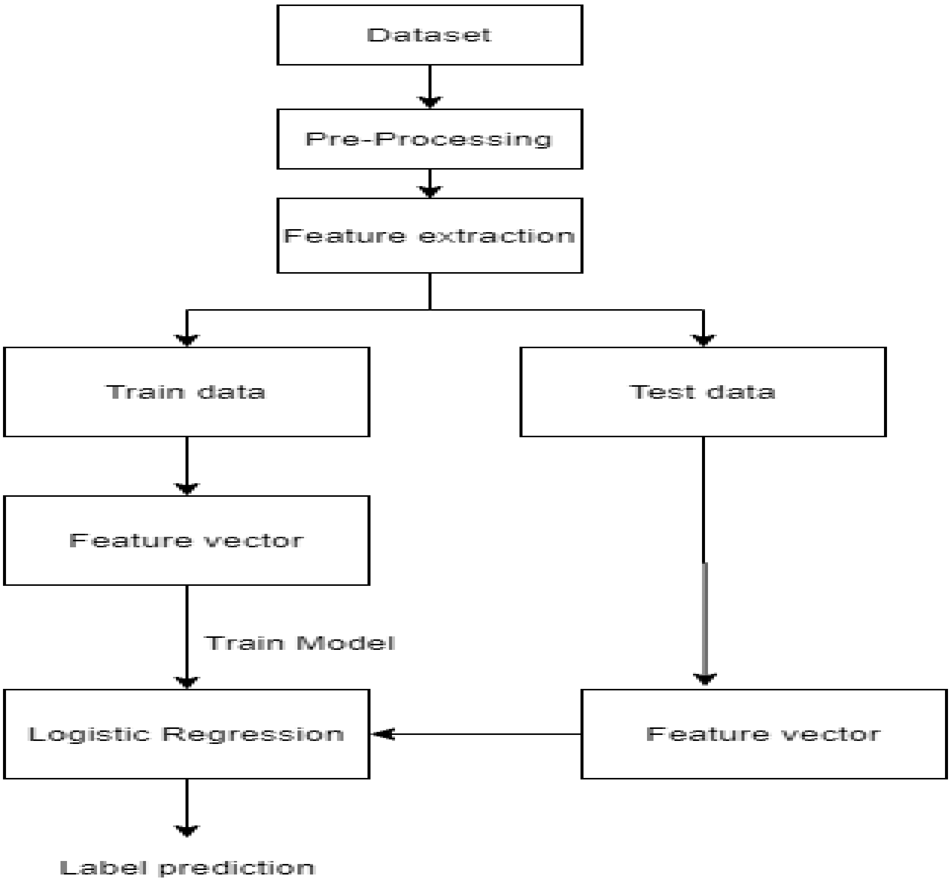
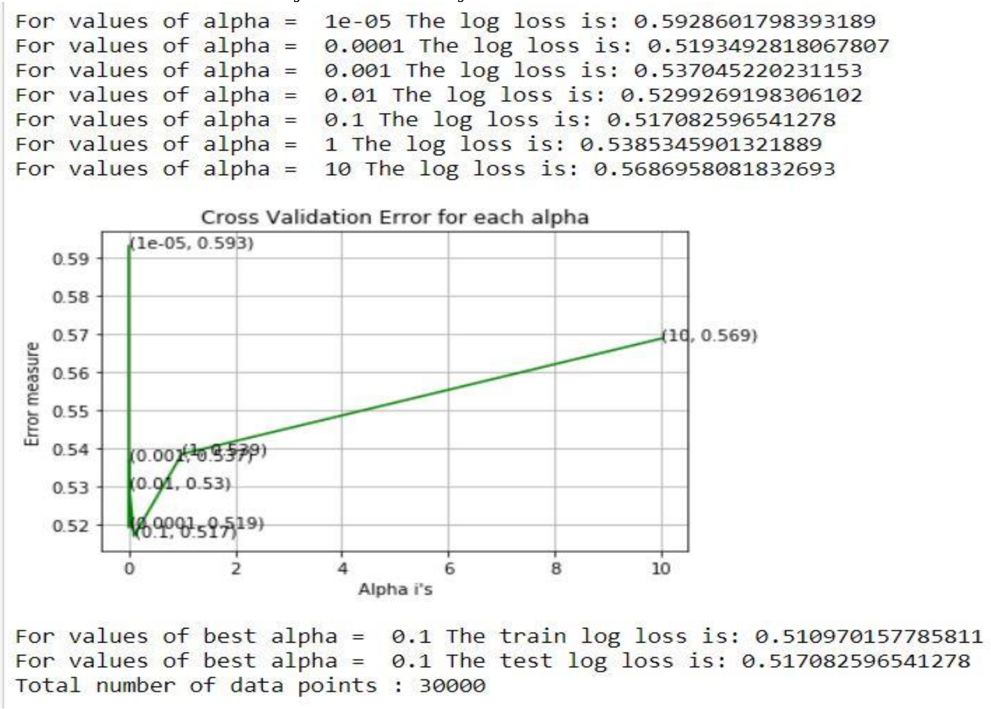


Fig 1- Model Workflow

First of all pre-processing is done to the dataset. After pre processing, feature extraction is done using NLP where certain attributes are calculated and added on to the dataset which helps to analyze the delicacy. Further the data set is divided into training set and test set. The training set comprises of 70% of data and test set comprises of 30% of data. The these are converted into computer understandable format and given as an input to model. The model is built using Logistic Regression on the training set and test set is used to analyze the accuracy of the model.



```
For values of alpha =  1e-05 The log loss is: 0.5928601798393189
For values of alpha =  0.0001 The log loss is: 0.5193492818067807
For values of alpha =  0.001 The log loss is: 0.537045220231153
For values of alpha =  0.01 The log loss is: 0.5299269198306102
For values of alpha =  0.1 The log loss is: 0.517082596541278
For values of alpha =  1 The log loss is: 0.5385345901321889
For values of alpha =  10 The log loss is: 0.5686958081832693
```

```
For values of best alpha =  0.1 The train log loss is: 0.510970157785811
For values of best alpha =  0.1 The test log loss is: 0.517082596541278
Total number of data points : 30000
```

Fig 2 Accuracy of the Model

The figure 2 depicts the log-loss value of Logistic regression model which is around 0.5109 at alpha 0.1. At this point of the the accuracy of the model is the best.

4. RESULTS

In this work, first of all data has been downloaded from data source, Quora and uploaded. The total number of data in the dataset is around 4 lacks sets of questions. Data pre-processing has been carried out, which basically involves like removal of stopwords, conversion of the text into lower cases, removing punctuations. Subsequently feature engineering is also performed on the cleansed dataset. Due to feature engineering few additional attributes are calculated such as q1len, q2len, word_common, word_Total, word_share etc. Further more advanced feature engineering is performed which adds few more attributes such as fuzz_ratio, token_ratio which helps us to analyze the probability of similarity between the words of two questions. Based on these features on a random model algorithm achieves a log-loss of 0.884268. By applying Logistic Regression the Log-loss comes down drastically to 0.5109 which shows that Accuracy of the model is much more efficient from random model.

5. REFERENCES:

[1] Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiský, Phil Blunsom. Reasoning about Entailment with Neural Attention. 2016. [Online]. Available: https://arxiv.org/abs/1509.06664

[2] Minh-Thang Luong, Hieu Pham, Christopher D. Manning. Effective Approaches to Attention-based Neural Machine Translation. 2015. [Online]. Available: https://arxiv.org/abs/1508.04025

[3] Jonas Mueller and Aditya Thyagarajan. Siamese Recurrent Architectures for Learning Sentences Similarity. 2016. [Online]. Available: https://dl.acm.org/doi/10.5555/3016100.3016291

[4] Shuohang Wang and Jing Jiang. Learning natural language inference with LSTM. In Proceedings of NAACL, 2016. [Online]. Available: https://www.aclweb.org/anthology/N16-1170/

[5] Samuel R. Bowman, Gabor Angeli, Christopher D. Manning. A large annotated corpus for learning natural language inference. In EMNLP,2015. [Online]. Available: https://arxiv.org/abs/1508.05326

[6] Minh-Thang Luong, Illya Sutskever, Quoc V. Le, Oriol Vinyals, Wojciech Zaremba. Addressing the Rare Word Problem in Neural Machine Translation. In ACL, 2015. [Online]. Available: https://arxiv.org/abs/1410.8206

[7] Minh-Thang Luong, Illya Sutskever, Quoc V. Le, Oriol Vinyals, Wojciech Zaremba. Addressing the Rare Word Problem in Neural Machine Translation. In ACL, 2015. [Online]. Available: https://arxiv.org/abs/1410.8206

[8] Illya Sutskever, Quoc V. Le and Quoc V Le. Sequence to sequence learning with neural network. In NIPS, 2014. [Online]. Available: https://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf

[9] Jeffrey Pennington, Richard Socher and Christopher D. Manning. 2014. GloVe:Global Vectors for Word Representation. [Online]. Available: https://www.aclweb.org/anthology/D14-1162.pdf

[10] Jiang Zhao, Tian Zhu and Man Lan. Enchu: One stone two birds:Ensemble of heterogenous measures for semantic relatedness and textual entailment. In SemEval, 2014. [Online]. Available: https://www.aclweb.org/anthology/S14-2044.pdf