Final Project Report

on

Message Classification for Twitter Data

IIT PATNA,

Patna, Bihar, India

2016-17

Submitted by,

Prasant Kapil

Roll No: 1721CS02

&

Sovan Kumar Sahoo

Roll No: 1721CS04

Table of Contents

Abstract

Every day millions of twitter user tweets their views on various topics using short messages of 140 characters length. Tweet classification is a process of classifying the tweets based on the topics using the keywords of the tweets as feature. This work is going to classify the tweets into six broad categories using deep neural network. The categories are Sports, Politics, Entertainment, Education, Technology and Business. Supervised learning techniques is used to classify the messages. The input to this programme is natural language text and this application will give the best possible class about the messages.

Keywords— Tweet classification, Deep neural network, Supervised learning

2. Introduction

This is the era of social media and tweeter is one of the most popular social media. Twitter is a microblogging site where users express view in the form of 140 character short messages which are called tweets. Everyday millions of tweets are generated. This huge data has introduced new avenues of research works which have both academic and business importance. Twitter message classification and twitter sentiment classification are the two main areas where many researcher are working around the world. Our work is on tweeter message classification using deep neural network.

2.1 Literature Survey

Twitter message classification is a hot topic in academia. Many researcher are working in this particular topic. A new tweet classification Method that makes use of tweet features like URL's in the tweet, retweeted tweets and influential users tweet is proposed by Dr. A. Suruliandi et. al. [1] Their experiments were carried out with extensive tweet data set which they manually collected from various trending topics and users. Then they compared performance of the proposed algorithm in classifying the tweets with the text classification algorithms like SVM, Naïve Bayes, KNN etc. Their proposed method outclasses the conventional text classification algorithms in classifying the tweets. Their proposed method made use of tweet features URL's in the tweets, Retweeted tweets and tweets from the most influential user of a trend topic as features for the tweet classification. If the tweet contains URL then the web page of that URL is categorized accordingly. Then the tweet is categorized in to the same category. If the tweet contains Trend topics, then words from top five retweeted tweets of that trend and top five tweets of the influential are collected and the collected word is classified using conventional text classifiers. If the tweet doesn't have URL or trend topic, then a conventional text classifier is used to classify the tweet.

Another work on Twitter Trending Topic Classification is done by Ramanathan Narayanan et al.[2]. They experimented with 2 approaches for topic classification; (i) the well-known Bag-of-Words approach for text classification and (ii) network-based classification. In text-based classification method, they construct word vectors with trending topic definition and tweets, and the commonly used tf-idf weights are used to classify the topics using a Naive Bayes Multinomial classifier. In network-based classification method, they identified top 5 similar topics for a given topic based on the number of common influential users. The categories of the similar topics and the number of common influential users between the given topic and its similar topics are used to classify the given topic using a C5.0 decision tree learner.

In [3] Twitter News Classification Using SVM is done by Inoshika Dilrukshi, et al. The purpose of this research is to classify news into different groups so that the user could identify the most popular news group in a given country for a given time. The short messages were extracted from Twitter micro blog. Several active news groups were chosen to extract the short messages. Each short message was classified manually into 12 groups. These classified data were used to train the machine learning techniques. Words of each short message was considered as features and a feature vector was created using bag-of-words approach in order to create the instances. The data were trained using SVM (Support Vector Machine) machine learning techniques.

3. Resources

Data Resource

We manually collected the tweets from various trend topic from www.twitter.com

Tools

- We have done this work in Python programming language. We have used PyCharm IDE (Community Edition)
- For training the neural network we have used Keras
- For Result analysis and report we have used Microsoft Office Packages

4. Work Done

4.1 Data Set

- **Description of the Data:**

The Data that are used were collected manually from six different fields which are Sports, Business, Technology, Entertainment, Politics and Education. Table 1 shows the distribution of messages.

Topics (Class)	Number of Tweets
Sports	303
Entertainment	204
Business	230
Politics	206
Education	152
Technology	167

Table 1: Diversity of tweet data set

- **Pre-processing of Data**

  The following pre-processing steps were done

  - Removal of Punctuation and Stop words
  - Remove the Word containing '#', '@' and 'http' or 'https' which means the trending topics, user name and URL are removed from tweets
  - Numerical character and special character are removed from the tweets
  - Each word of the tweets are lemmatized

- **Data Representation**

We have used bag-of-word technique for data representation. The bag-of-words model is a simplifying representation used in natural language processing. In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity.

4.2 Final Architecture

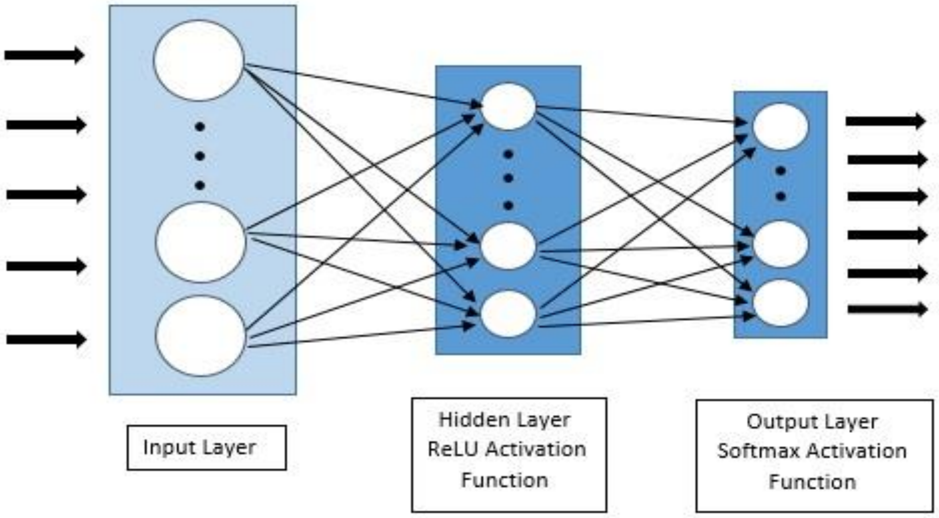

Figure 1: Neural Network Architecture

For training we have used a network topology of simple one-layer neural network. Figure 1 depicts the diagram of the network model. We have used Rectified Linear Unit (ReLU) as activation function. In the output layer we have used Softmax activation function for the probabilistic approximation of the output. The brief description of ReLU and Softmax is given below.

ReLU: The Rectified Linear Unit has become very popular in the last few years. It computes the function f(x) =max (0, x). In other words, the activation is simply thresholded at zero. There are several pros for using the ReLUs:

- It was found to greatly accelerate the convergence of stochastic gradient descent compared to the sigmoid/tanh functions. It is argued that this is due to its linear, non-saturating form.
- Compared to tanh/sigmoid neurons that involve expensive operations (exponentials, etc.), the ReLU can be implemented by simply thresholding a matrix of activations at zero.

Softmax: Softmax function is used for the output layer only (at least in most cases) to ensure that the sum of the components of output vector is equal to 1. This also implies what is the probability of occurrence of each component (class) of the output and hence sum of the probabilities (or output components) is equal to 1

4.3 Results and Observation

- **Accuracy Plot for different Training and Validation sets data division**

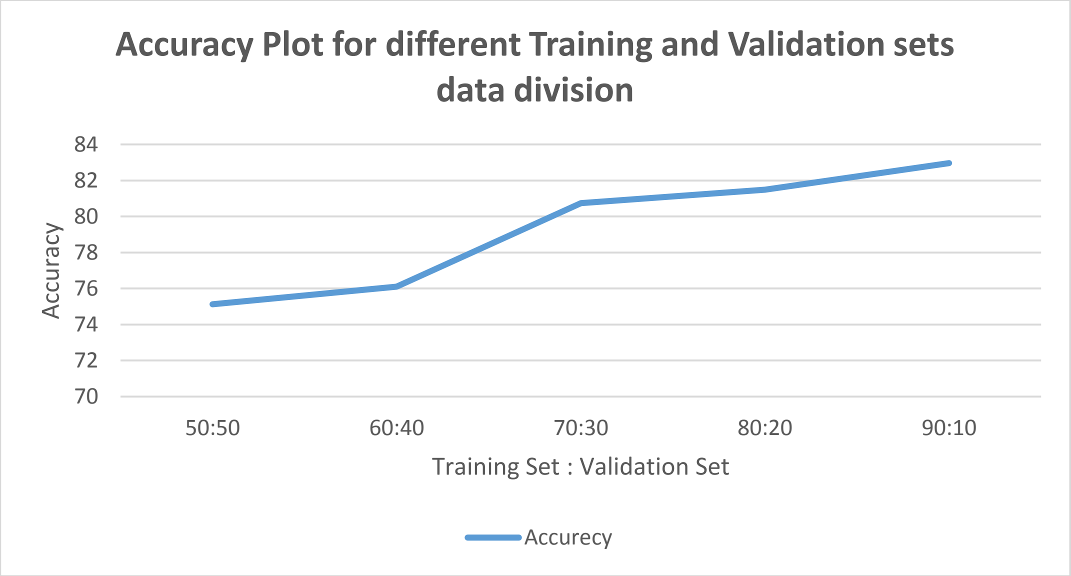

Figure 2

- **Observations**

From this experiment we have achieved almost 80% accuracy with 70:30 Training and validation dataset ratio. It is also observed that if we increase the size of training dataset then the accuracy is increasing. Batch size that we have used is 32 and number of epoch is 10.

4.4 Github location for your code

https://github.com/imprasshant/Twitter-Message-Classification

4.5 Future Work

In this work we have used small data set for training as well testing. Future work in this project include usage of larger data set for training and validation. Exploration of other neural network like RNN, CNN, LSTM in order to increase the accuracy. Inclusion of more topics and sub topics.

5. References

1. P.Selvaperumal and Dr.A.Suruliandi "A Short Message Classification Algorithm For Tweet Classification", 2014 International Conference on Recent Trends in Information Technology

2. Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md. Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary," Twitter Trending Topic Classification", 2011 11th IEEE International Conference on Data Mining Workshops
3. Inoshika Dilrukshi, Kasun De Zoysa, Amitha Caldera," Twitter News Classification Using SVM, The 8th International Conference on Computer Science & Education (ICCSE 2013) April 26-28, 2013. Colombo, Sri Lanka'

4. https://www.tensorflow.org/tutorials/word2vec

5. http://www.deeplearningbook.org/