International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307

Predicting the Best Tweet Using Machine Learning

Mr. L. A. Saleem[1] , I. Yagna Likhitha[2], E. Haritha[3], P. Jyothsna[4]
[1]Assistant Professor, Department of Computer Science & Engineering, BVRIT Hyderabad College of Engineering for women, Hyderabad, Telangana, India
[2-4]Department of Computer Science & Engineering, BVRIT Hyderabad College of Engineering for women, Hyderabad, Telangana, India

ABSTRACT

Twitter is an online social networking service that enables users to post and interact with messages known as tweets which are limited to 140 characters. Most of the communication here is done through following or retweeting. Retweet, is a main way to spread information in twitter. It is changing the geography of communication. Recently research focuses on analyzing the factors of retweet behavior. It's data feed includes all kinds of meta-data. In order to predict which tweet would be retweeted there are two ways. One is through studying the information propagation[1] path topology to build prediction model, but it is a very difficult task to construct the topology of user networks[2]. The other way is to build prediction model based on machine learning algorithm[3].Previously, a machine learning model was built using Theil-Sen Estimator[4] and a best fit line was fit and words in a tweet were scored. We built a machine learning model which not only scores the words in a tweet but also predicts[5] the best tweet among two tweets given as the input from the user. A basic fact is that different people are interested in different kinds of tweets, and they will retweet tweets which they are interested in. First, we collect tweets of different categories from valid account of famous news media as learning corpus which acts as a dataset for our machine learning algorithm.

Keywords: AI, Machine learning model, Theil-Sen Estimator, best fit line, retweet, score, topology, information propagation

I. INTRODUCTION

Why Machine Learning?

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve. Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

Machine learning algorithms are often categorized as supervised or unsupervised.

Supervised learning, which trains algorithms based on example input and output data that is labeled by humans, and unsupervised learning which provides the algorithm with no labeled data in order to allow it to find structure within its input data.

Dataset:

Almost thirty thousand tweets are collected from top 150 users of twitter which sums upto thousand tweets per user. This dataset is uses as both training and testing data. Main parameters are retweet count, date and time of each tweet. These are the dominant parameters which make makes the tweet likely to get retweeted.

Data Preprocessing:

After collecting the data it is preprocessed which includes cleaning[6], stemming[7], tokenization[8] and vectorization[9].

Data cleaning includes removal of unimportant data like hashtags, punctuation, emotions etc.

Stemming includes extraction of stopwords from the words in a tweet. Porter Stemmer algorithm is used for stemming as it has the highest accuracy(31.9% to 34.7%) among stemming algorithms.

Tokenization includes splitting the tweet with respective delimiters.

Vectorization includes conversion of text data to numbers used to plot the data.

System Model:

System Architecture:

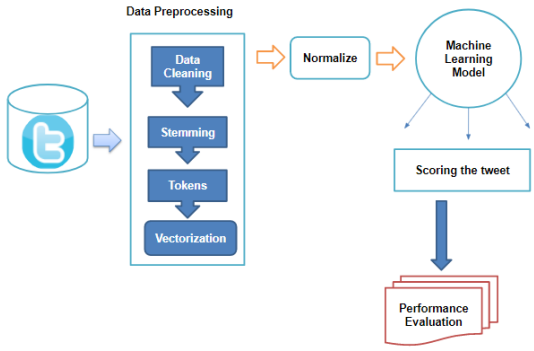
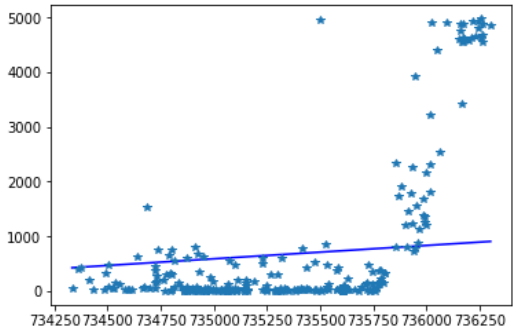


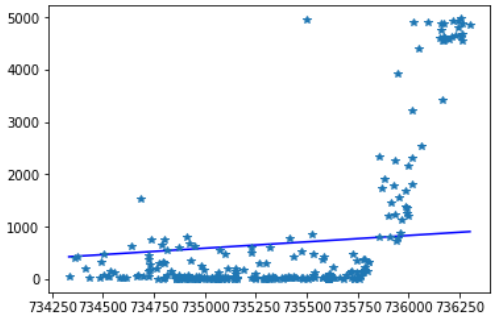
Figure 1

Scoring of Dataset

To score the data first we need to construct Theil-Sen estimator for each user's tweet history. Compared to the all other estimators, the Theil-Sen estimator is robust against outliers. It has a breakdown point of about 29.3% in case of a simple linear regression[10] which means that it can tolerate arbitrary corrupted data (outliers) of up to 29.3% in the two-dimensional case.



User1 tweets with Theil-Sen estimator



User2 tweets with Theil-Sen estimator

This regression line represents the increase in retweets of user which increases over a period of time. We discard all the tweets below regression line and score each tweet by its retweet count percentage above Theil-sen baseline. Score of each tweet is calculated as

Score of each tweet = (predicted value-actual value)/predicted value * 100

Actual value = actual retweet count of tweet

Predicted value = predicted retweet count of tweet by the model

For example if a tweet gets retweeted 40,000 times where as model predicts it as 20,000 its score would be

Score = (40,000-20,000)/20,000 * 100

Score = 100

I. LEARNING ALGORITHM

Construction Of Dictionary

A tremendous factor for concluding the quality of tweet is content present in the tweet. We need to determine how powerful each word is in making the tweet retweetable. More powerful the words present in the tweet, more likely to get retweeted. In order to extract the content of the tweet we construct one dictionary from the given training data. Our algorithm assigns a tweet score to each and every word present in the dictionary

Scoring Of Words:

Calculation of score for each and every word is done by looking at each tweet present in the dataset.

Algorithm sums up all the tweet scores of the tweets in which that particular word(for which we need to calculate the score) is present. Finally it averages the score and assigns that score to the word in the dictionary. In this ways our algorithm calculates scores for all the unique words present in the dictionary.

For example calculating score of SALMON from the following tweets.

Tweet 1: I've seen tastier canned salmon → 12

Tweet 2: I like smoked salmon → 6

Word Score = (6+12)/2 => 9

Prediction of score for the tweet:

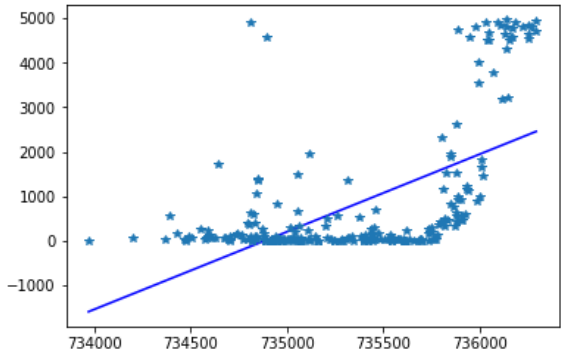
We assign a score for each and every word present in the new tweet by applying dictionary of word scores. Final score for the tweet is calculated by averaging all the scores of the words present in the tweet.

For example : "Smoked salmon today"

(78+9+21)/3 => 36

II. LINEAR REGRESSION

We built a best fit line using linear regression between time and the the score of the tweet to understand the increase in the score of the tweet over a period of time.



III. RESULTS

Using our learning algorithm for back-end , we built an application which takes two tweets as input from the user and predicts which tweet thereby has the highest possibility of getting retweeted.

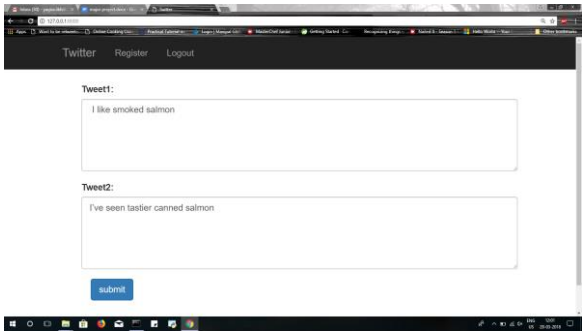


Figure 2

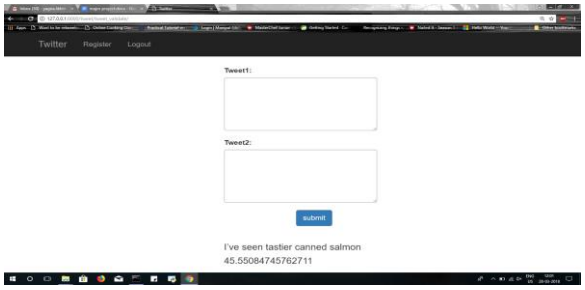


Figure 3

IV. CONCLUSION

Our learning algorithm was able to predict the best tweet which is more likely to get retweeted between two tweets given by the user. Apart from the existing tweets, it was able to predict the score of new tweets, which conclude that our algorithm did learn something from the dataset. The efficiency of the learning algorithm depends on the dataset and its preprocessing. As we preprocessed the data by cleaning, stemming, tokenization and vectorization it

is efficient. Also, as we added few identified regular expressions to decontract a word like I'll to I will which in turn increases the efficiency of the learning algorithm. The concept of this scoring and predicting the best tweet maximizes the advertising campaigns on twitter[11].

V. FUTURE WORK

The learning algorithm can be made better by increasing the size of the dataset with more tweets especially different categories of tweets for better prediction. Also, future work could include processing hashtags(#hashtags) and emotions which generally add to the direction of the tweet. Scoring can also be improved by processing pairs of words using semantic computation.

VI. REFERENCES

[1]. Sadikov,E. & Martinez,M. "Information Propagation on Twitter." CS322 Project Report 2009.

[2]. Uysal,I. and Croft,W. B. 2011. User oriented tweet ranking: a filtering approach to microblogs. In proceedings of the 20th ACM international conference on Information and knowledge management (CIKM '11),Bettina Berendt,Arjen de Vries,Wenfei Fan,Craig Macdonald,Iadh Ounis,and Ian Ruthven (Eds.). ACM,New York,NY,USA,2261-2264. DOI=10.1145/2063576.2063941

[3]. Scikit-learn: Machine Learning in Python,Pedregosa et al.,JMLR 12,2011,pp. 2825-2830.

[4]. Dang Xin,H Peng,X Wang and H Zhang (2008),Theil-Sen Estimators in a Multiple Linear Regression Model. Submitted paper.

[5]. L. Madlberger and A. Almansour,"Predictions based on Twitter-A critical view on the research process," 2014 International Conference on Data and Software Engineering (ICODSE),Bandung,2014,pp. 1-6.doi: 10.1109/ICODSE.2014.7062667

[6]. Yoon S,Elhadad N,Bakken S. A Practical Approach for Content Mining of Tweets. American journal of preventive medicine. 2013;45(1):122-129. doi:10.1016/j.amepre.2013.02.025.

[7]. Anjali Ganesh Jivani ,A Comparative Study of Stemming Algorithms,International Journal of Computer,Technology and Application,Volume 2,ISSN:2229-6093.

[8]. S. Diaz-Santiago,L. M. Rodriguez-Henriquez,D. Chakraborty,"A cryptographic study of tokenization systems",International Journal of Information Security,vol. 15,no. 4,pp. 413-432,2016.

[9]. K. Lang,"20 newsgroup data set". Available at: qwone.com/~jason/20Newsgroups/. Accessed 30-Sep-2015

[10]. Srivastava,A.K. and Shalabh (1995): "Predictions in Linear Regression Models With Measurement Errors",Indian Journal of Applied Economics,Vol. 4,No. 2,pp. 1-14.

[11]. Y. Mei,W. Zhao and J. Yang,"Maximizing the Effectiveness of Advertising Campaigns on Twitter," 2017 IEEE International Congress on Big Data (BigData Congress), Honolulu, HI,2017, pp. 73-80.doi: 10.1109/BigDataCongress.2017.19