Predicting the Impact of Selected Quotes On Social Media Counts

by

Priya Mukherjee and Komal Thakkar

Department of Computer Science, Northern Illinois University

ABSTRACT:

The advancement in the social media technology has opened broad avenues of digital marketing. It is a great platform where the users can promote their work, blogs and their business. Twitter is one of the best social media sites where the user can get instant reaction about their work. In this paper, we investigated the impact of the interesting snippet of the conversation or more precisely the post on the twitter posts count. Whenever the researcher or the user post about their paper in any social media sites, it is seen that mostly there is a brief information about the paper with the link of the paper. It is hard for the user to open the link and read the paper and then express their opinion about it. By this quote we are predicting do selected quotes grab any attention from the user to improve the twitter post counts. In this endeavor, we built regression models to achieve this purpose based on the mean accuracy.

KEYWORDS:

Altmetrics, Selected Quotes, Social media, Twitter Counts, Bag of Words, Natural Language Processing

1 BACKGROUND:

1.1 Introduction:

With technology developing as it is, and social media being the most popular form of online communication for most people, there is certainly no shortage of platforms in which to share, explore and communicate. Social media have become important complementary channels for disseminating and discovering research. It is one of the most active platforms in research lifecycle [1]. Twitter is one of the most user-friendly platforms where users interact with one another to discuss about the situations that motivate people to share their thoughts publicly [5]. The relationship exists between the scientific text and the public understanding of the text [2] which means somehow the text impact the posts count. Research evaluations becomes broader as the societal products (outputs), societal use (societal references), and societal benefits (changes in society) of research come into scope [3][4].

Users interaction in twitter platform provides footprints to monitor the expression of the users towards other's work. Researchers can trace how users like their work by the number

of times their paper has been shared or retweeted.

1.2 Data Collection:

The data used in this study was obtained from Altmetric.com. Altmetrics data is varied metrics which support robust research into usable impact measures [7]. It collects all the necessary information about the research articles [8]. Altmetrics plays a vital role to predict the relationship between the research paper and social media [6]. Our goal is to predict the impact of the selected quotes in Twitter. Thus, while extracting the data we made sure that the quotes are available in Twitter. There was a situation where each article has more than one quotes. We handled it by concatenating all the quotes into one quote which refer to each article has only one quote. Altmetrics contains very huge number of scientific articles. We have extracted the data by random sampling which resulted in total 15,000 of data.

1.3 Text Cleaning:

As we are dealing with text data, it is important that we preprocess or clean the data. Regular expressions have been used to remove URLs, special characters, whitespaces and digits. Additionally, NLTK library has been imported to filter the most frequent words like 'the', 'is', 'a', and so on. This is achieved by using stop words method to the text. While analyzing the data, we found that there were quotes that was in different language other than English. Used NLTK corpus library to filter out non-English words by checking the word in the English dictionary. If the word is not present in the dictionary, it is filtered out from the dataset.

2 METHODS:

2.1 Feature Generation:

Our features are all the distinct words we extracted from our quotes. Bag of Words model achieves this.

Bag of Words:

Bag of words model is one of a series of techniques from a field of computer science known as Natural Language Processing or NLP to extract features from text. The way it does this is by using Count Vectorizer. It converts a collection of text documents to a matrix of token counts. These counts are the frequency of words in a document.

creationism	evolution	big	bang	classroom	infinite	two	dimensional	system
1	1	1	1	1	0	0	0	0
0	0	0	0	0	2	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	4	0	0	0

Figure 1: Feature extraction using Bag of Words.

In Figure 1, each row is a selected quote (here the quote is after concatenation of multiple quotes into one quote per article) from different articles. Each word present in the column represent the feature of the data. For the first selected quote, all the words left after text cleaning are added as a feature with its frequency (i.e., how many times a word occurred in the quote). While processing the second quote, it will first check whether all the words are distinct from the words already present as a feature. If all the words are

different then all the words (vectors) will be appended to the feature. If there are some words that already present in the feature then it will increase it frequency by the number of times it appeared in the second quote and rest words will be added to the feature. This process will continue until the last selected quote is processed. Total number of words (features) extract from the selected quote is 13857. Figure 2 show the most popular top ten words used in posts (selected quotes) while posting about a research article.

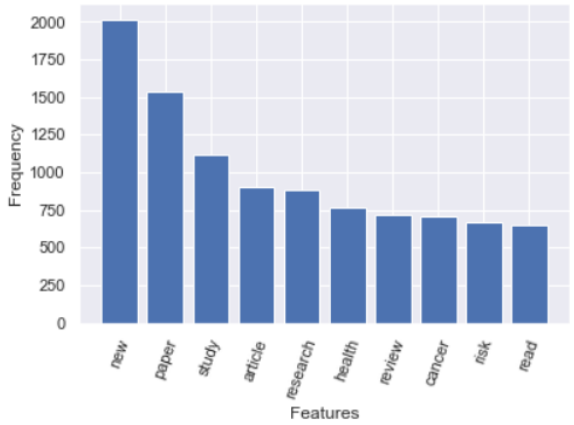

Figure 2: Top 10 words (features) present in the selected quote.

2.2 Target:

In this paper we are predicting the impact of selected quotes only on twitter count, we classified the class label based on the mean of the twitter counts. As twitter counts are the random numbers, if we would have considered it as our target, then it would have led to an imbalanced dataset. Thus, we decided to classify it based on the average of the counts. If the number of post count is greater than the mean of itself, then the class label is 'a' and if the number of post count is greater than its own mean then the class is labelled as 'b'.

2.3 Principal Component Analysis (PCA):

2.3.1 Speed-up Machine Learning Algorithms:

Principal Component Analysis (PCA) is a simple yet popular and useful linear transformation technique that is used in numerous applications. The number of features used in our project is quite huge. Our learning algorithm is too slow because the input dimension is too high. Thus, we need some technique to speed up the fitting of a machine learning algorithm by changing the optimization algorithm. A more common way of speeding up a machine learning algorithm is by using Principal Component Analysis (PCA). It reduces the dimensionality by combining the features. We split the data into 70% training and 30% test data. Our database of handwritten digits is more suitable as it has 13857 feature columns (13857 dimensions), a training set of 10,500 and a test set of 4,500.

2.3.2 Standardize the Data:

PCA is affected by scale so you need to scale the features in the data before applying PCA. You can transform the data onto unit scale (mean = 0 and variance = 1) which is a requirement for the optimal performance of many machine learning algorithms. StandardScaler helps standardize the dataset's features. standardize the dataset's features. We fit on the training set and transform on the training and test set.

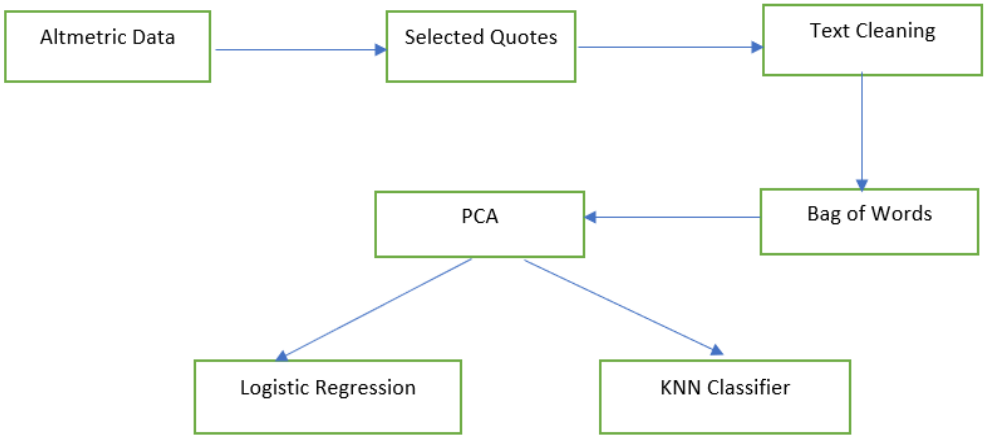

Figure 3: Pipeline for prediction of selected quote impact

2.3.3 Apply PCA:

During this process, it fits the training set based on the variance to be retained. In our project, we provided the variance as 90% which means that Scikit-learn choose the minimum number of principal components such that 90% of the variance is retained. In other words, we provide the variance so that we don't lose the useful data. After this, fit PCA on training set. Using pca.n_components_ we found that only 4881 features from 13857 have been selected by PCA.

3. Results and Discussion:

3.1 Logistic Regression:

We applied Logistic Regression to predict the mean accuracy on the test data. 'lbfgs' solver has been used to optimize the code.

3.2 KNN (K Nearest Neighbor) Classifier

After Logistic Regression classifier, we also checked the score (mean accuracy) of our data with KNN classifier. We got a high accuracy of 0.96. Table 1 represent the score predicted from both Logistic Regression and KNN classifier. We also performed these algorithms on data without PCA application. But it resulted in no difference in the score.

Methods	**Accuracy**
Logistic Regression	0.95
KNN Classifier	0.96

Table 1: Mean accuracy of Logistic Regression and KNN Classifier.

4. Conclusion and Future Work

In this paper, we investigated the impact of selected quotes on twitter post counts. Figure 3 gives the overview of our learning model. As the words were our features, we processed the text and used Bag of Words model which is a technique in Natural Language Processing (NLP) to extract the features from the selected quotes. The features are the distinct words which we fetched from the quotes. We classified the target based on the mean of the twitter posts count. As the size of our dataset is huge, we applied PCA

algorithm to speed up our learning model. Predicted the impact via regression classifier and KNN classifier. The mean score (accuracy) were almost equal with both the methods.

In future, we plan to work with the entire dataset to find the relations of the quotes with other social media sites. Also, plan to analyze the sentiment of the quotes and public understanding. We can apply our research to analyze the current hot topics in particular field like 'cancer' in medicine.

5. References

1) Rowlands, D. Nicholas, B. Russell, N. Canty, and A. Watkinson, "Social media use in the research workflow," Learn. Publ., vol. 24, no. 3, pp. 183–195, Jul. 2011.
2) Harish Varma Siravuri, Akhil Pandey Akella, Christian Bailey, and HamedAlhoori. 2018. "Using Social Media and Scholarly Text to Predict Public Understanding of Science". In JCDL '18: The 18th ACM/IEEE Joint Conference on Digital Libraries, June 3–7, 2018, Fort Worth, TX, USA. ACM, New York NY, USA, Article 4, 2 pages.
3) L. Bornmann, "What is societal impact of research and how can it be assessed? a literature survey," J. Am. Soc. Inf. Sci. Technol., vol. 64, no. 2, pp. 217–233, Feb. 2013.
4) R. Smith, "Measuring the social impact of research," BMJ, vol. 323, no. 7312, p. 528, Sep. 2001.
5) Axel Bruns, Katrin Weller, (2014) "Twitter data analytics – or: the pleasures and perils of studying Twitter", Aslib Journal of Information Management, Vol. 66 Issue: 3
6) J. Priem, H. A. Piwowar, and B. M. Hemminger, "Altmetrics in the wild: Using social media to explore scholarly impact," arXiv:1203.4745, Mar. 2012.
7) E. Adie and W. Roe, "Altmetric: enriching scholarly content with article-level discussion and metrics," Learn. Publ., vol. 26, no. 1, pp. 11–17, Jan. 2013.
8) J. Priem, P. Groth, and D. Taraborelli, "The altmetrics collection," PLoS One, vol. 7, no. 11, p. e48753, Nov. 2012.