2022

Predicting Brand Sentiment on Twitter

2020-04-04

PUSHPENDRA SHARMA | BRAINSTATION | TORONTO

1. Problem Statement

The goal of this project is to help business owners predict real time public sentiment about a brand on twitter. This project utilizes tweets from twitter, natural language processing techniques, machine learning models, and deep learning models to predict the sentiment of tweets towards a brand over a specified time interval. This study will help companies that has online social media presence in understanding customer behavior, early crisis detection, campaign performance analysis and competitive analysis.

2. Background

4.62 billion people that is half of the world (58.4%) now use social media to connect with their family and friends [1]. The average daily time spent using social media is 2h 27m [1]. Although, social media made it easy for companies to target potential customers, but it also gave customers a platform to provide feedbacks on brands. According to a study by Nielsen [2], 92% of users trust online content from friends and family above all other forms of brand messaging. Further, 53% of millennials have said that user-generated content has influenced their purchasing decision. These data show the importance of maintaining a positive public brand outlook on social media platforms to keep customers happy.

3. Data Source

The data for this project was acquired from the sentiment 140 website [3]. This data was originally scrapped by Go et al. (2009) [4] using the twitter's API. This dataset contains 1.6 million tweets for multiple users from April 2009 to June, 2009. The tweets have been annotated with 3 sentiments (0 = negative, 2 = neutral, 4 = positive) based on emoticons in the tweets. The authors have annotated a tweet with a positive sentiment if it contains positive emoticons like, :). Similarly, the tweet was annotated as negative if it contains negative emoticons such as :(. The tweets were scrapped from twitter using specific keywords in twitter's API. These tweets are used to train machine learning and deep learning models to predict sentiment about a brand. Additionally, in order to test the developed models, twitter's API v2 [5] was used to scrape tweets for a specific brand.

4. Data Processing

The original data contained five columns as shown in *Table 1*. Each of the columns were explored individually to understand their relationship with target variable. For modelling purpose, the text of tweet was used as a feature to predict the target (sentiment of tweet). It was observed that the text in tweet contained usernames and links which were removed form the text. The columns id, date and user were dropped from the data for modelling purposes.

Table 1: Snapshot of data - different columns and their values.

Id	Date	User	Tweet	Sentiment
2071344601	2009-06-07	Redrockinrory	what the heck I smell like camp fire and I don't want to get up early	Negative (0)
1996174719	2009-06-01	vacant_heart	nice name do start posting soon	Positive (1)

5. Exploratory Data Analysis (EDA)

Figure 1 shows the distribution of target variable that is the sentiment behind a tweet. The target variable had approximately same proportion of tweets with the positive and negative sentiment.

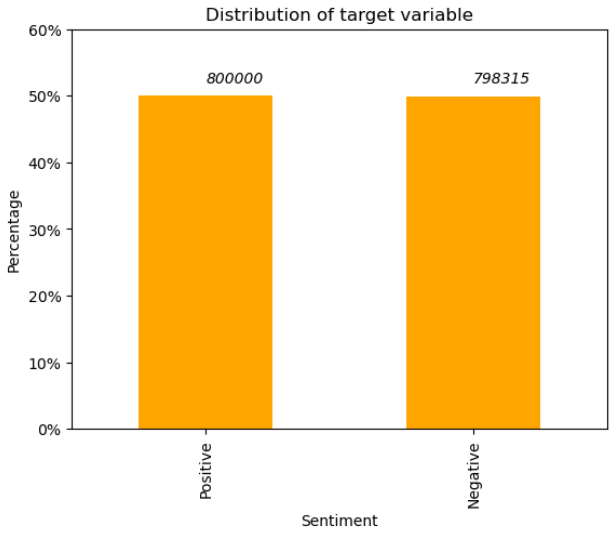

Figure 1: Distribution of target variable (sentiment of tweet)

The number of tweets in the data are highest for the month of June followed by May and April. Although tweets were scrapped only for 19 days in the month of June, the number of tweets were still greater than the number of tweets for the month of May.

6. Natural Language Processing and Modelling

Stemming was performed on the text data after removing stop words to get the root English words that would be predictor of positive or negative sentiment. Bag of words and TF-IDF vectorizers were used to transform text data to machine readable tabular numeric data.

Machine learning models such as Logistic Regression, XGBoost, Naïve Bayes, and Support Vector Classifier were fitted on the transformed data. Deep learning models such as neural

networks and word2vec embeddings were also fitted on the training data to predict sentiment of tweets. Validation set was used for hyper parameter tuning. Accuracy of some of the machine learning and deep learning models are provided in *Table 2*.

Table 2: Accuracy on the validation sets for best fitted models.

Model	Vectorizer	Accuracy
Logistic Regression	Countvectorizer and Stemming	0.77
XGBoost	Countvectorizer and Stemming	0.75
Neural Network	TF-IDF and Stemming	0.72
Neural Network with word embeddings	Simple preprocess and Stemming	0.75

It should be noted that neural networks were trained only on the 10% sample of data for memory management and computational efficacy.

7. Findings

Figure 2 shows the coefficients from logistic regression model for top 20 features that are best predictors of positive and negative sentiment for a tweet.

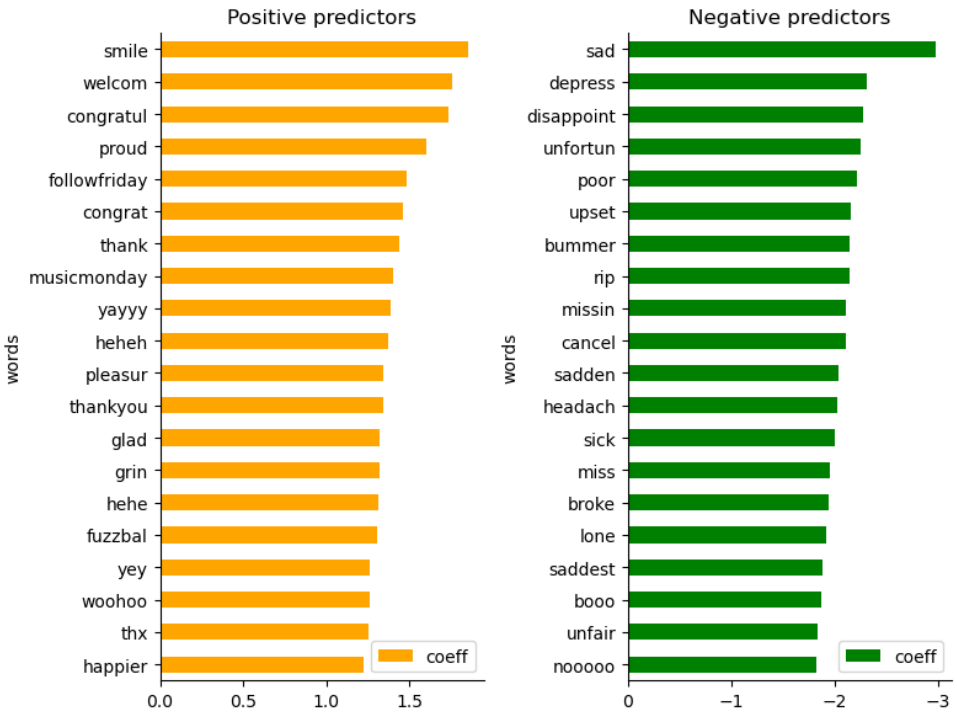

Figure 2: Top 20 predictors of positive and negative sentiment for a tweet.

The top 20 predictors for negative and positive sentiment for a tweet are self explanatory and intuitive. We know that the words such as smile, welcome, congratulations, proud carry a

positive sentiment with them while the words sad, depress, disappoint, unfortunate, poor carry a negative sentiment with them. The word smile was top predictor of positive sentiment while the word sad was top predictor of negative sentiment.

The sentiment for a brand was predicted based on the sentiment of tweet that contained the brand name. To test the models and predict sentiment, tweets with tag '#Nike' were scrapped from twitter. *Figure 3* shows the change in sentiment over time for Nike brand. We can see that the sentiment score is above 0.5 which indicates that people are talking positively about the brand.

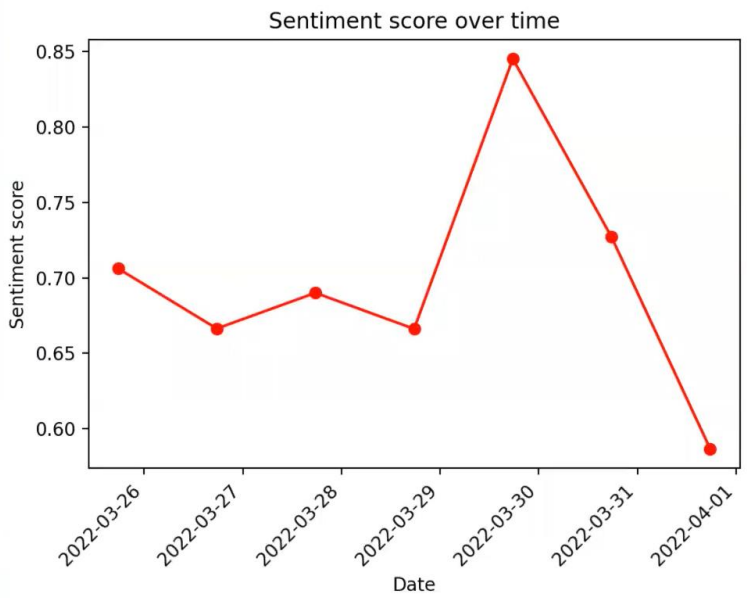

Figure 3: Variation of sentiment score over last seven days for Nike.

8. Conclusion

In this study, we used natural language processing coupled with machine learning and deep learning models to predict sentiment of tweets. The best fitted logistic regression model had an accuracy of 0.77 and performed better than other machine learning models. In order to predict public sentiment towards a brand, tweets were scrapped from twitter using twitter's API. Finally, an online app was developed to predict real time sentiment for a given brand that can help in understanding customer behaviour, early crisis detection and competitive analysis.

9. Next Steps

Neural networks were only trained only 10% sample of the data due to limited computational resources and yet they had an accuracy of 0.75. A next step would be to acquire more computational resources (cloud computing) to train neural networks for improving accuracy of the models. Further, the order of sequence in the text was not taken into account while

modelling, the future studies will use sequential models such as Recurrent neural networks to improve model accuracy. Finally, currently the app only calculates sentiment score based on Logistic regression model, the future work will improve app to predict sentiment score based on user provided models.

10. References

1. https://brainstation.io/course/online/remote-data-science-bootcamp
2. https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media
3. https://www.nielsen.com/us/en/insights/report/2012/global-trust-in-advertising-and-brand-messages-2/
4. http://help.sentiment140.com/for-students
5. https://www-cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf
6. https://developer.twitter.com/en/docs/twitter-api