Sentiment Analysis using Social Media

Grand Text Auto

1. INTRODUCTION

Social Media has become a ubiquitous source of information in the modern world. Many people use multiple platforms for social media, consuming for both entertainment and professional purposes. From a marketing perspective, it has become impossible to ignore the influence that Social Media has on consumers. Social Media also provides companies with the ability to directly interact with their consumers, gauging their interest in products and services. As automated techniques for text analysis grow more popular, it is now possible to understand the underlying sentiments of consumers. Latent thoughts and perspectives can be detected using techniques such as Sentiment Analysis.

In this project, we will explore the effect of news and communications from brands on the overall sentiment about the respective products[5]. Specifically, we will focus on Sony Playstation and Microsoft Xbox brands with both planning to release their computing products in November 2020. We plan to use sentiment analysis to find out what terms were positively associated with each product and what terms were negatively associated. This type of analysis can be useful for a company to find out where they are trailing their competition and use that knowledge to plan future communications.

To achieve our objective, we will train a model based on previously labeled data and use it to predict sentiment on new tweet data related to Xbox and Playstation. We will use these predicted labels to analyze the change in consumer sentiment towards these products as they approach their release.

2. DATASET DESCRIPTION

For Training our models, we are using the sentiment140 dataset from Kaggle[6]. It contains 1.6m tweets that have been labeled for the sentiment. The labels range between 0 and 4, with 0 being negative sentiment, 2 being neutral sentiment, and 4 being positive sentiment.

The dataset contains the following attributes for each tweet:

- The **target** sentiment
- The **id** of the tweet
- The **date** in which the tweet was published
- The **user** that sent the tweet
- The **content** of the tweet

Using the Twitter API Tweepy[2], we will scrape tweets related to the Xbox and Playstation as the test dataset, which are all unlabeled texts. We will use this data to evaluate the change in sentiment over time for the two products.

3. UPDATED PROGRESS SINCE PHASE 1

We have gotten all the data we need for training our model, which is the Sentiment140, and some test data using Tweepy. We noticed that although Tweepy has a rate limit of 300 calls/ 15 minutes, which is more than sufficient for our purposes, it only allows access to a tweet's data if you know the tweet id. Therefore, we need a workaround to get the tweet ids. For this, we will use the snscrape library[1].

The snscrape library uses the twitter search function to get the tweet ids. For example, the following query will fetch the tweet links for 300 tweets from 1st to 2nd September that match our target search words. These links are then stored in a file and can be used by Tweepy to get the tweet ids.

snscrape –max-results 300 twitter-search "(playstation 5 OR xbox series x OR ps5 OR xsx OR xss) lang:en until:2020-09-02 since:2020-09-01" > tweets.txt

The data retrieved for the tweet is as follows:

- id : tweet id
- username : user name of tweet sender
- retweetcount : number of retweets on the tweet
- text : the content of the tweet
- tweetcreateddts : date on which the tweet was created
- likes : number of likes received by the tweet
- hashtags : hashtags associated with the tweet
- followers : number of followers the user has
- location : location from which the tweet was sent

Our goal is to use the retweet count, likes, and twitter followers count to reduce duplication and a means to focus on the tweets that gained the most traction.

4. POSSIBLE APPROACHES

The authors in [7] discuss the major components of Sentiment Analysis pipelines and presents different baseline possibilities for adaptation. The authors discuss the three levels of sentiment analysis: document, sentence, and aspect/entity. A sentiment analysis pipeline is first to preprocess data, then preprocess the text, build classifiers, and then finally evaluate the classifiers. We will be extending this further by

using our classifier on new data for time-series evaluation. The authors note the reasons social media data (specifically tweets) are essential to analyze. The shortness of tweets (280 chars), online slang, hashtags, and twitter's diverse userbase make it interesting to study. They then detail the preprocessing steps that are important for Analysis: Remove retweets, social media related stop-words, punctuation. They also employ stemming and lemmatization in the preprocessing pipeline. Supervised learning techniques in the survey include SVM, Naive Bayes, Decision Trees, and Neural Networks. The authors also present unsupervised learning methods that involve using a known corpus of words with sentiment and providing positive/negative scores.

In [3], novel adaptations on top of normal preprocessing to augment the data is presented. The authors propose three different augmentations to feature vectors generated from tweets before training. The added features are total scores based on lexicon analysis of the text. The lexicon is used to compare tweets with known positive and negative sentiment words and add the total scores to each extracted feature vector. The authors found that this increased accuracy of prediction.

Since we are using the Sentiment140 as our training dataset, it is essential to understand how these tweets being analyzed and classified. Thus, we also read the research paper associated with this dataset [4]. The authors use distant supervised learning to deal with this task. In specific, they use a multinomial Naive Bayes model, a Maximum Entropy model using Standford Classifier, and the SVM^{light}, a linear kernel SVM software.

Another study uses a linear kernel SVM classifier is [8]. This classifier relies on features drawn from the training instances, and those obtained using external resources, including word and character N-grams, sentiment lexicon, target detection, POS tag detection, and encoding detection, has an F-score of 70.3. The study also derives 100-dimensional word vectors using the Word2Vec Skip-gram model trained over the domain dataset. The words embedded in a given tweet are taken to be the component-wise averages of the word vectors for all words appearing in the tweet. Using word vectors as an assist technique may be helpful since it can be used further in other complex approaches.

5. LIST OF DELIVERABLES

- Historical and Newest dataset collected via Twitter API
- Train models based on Sentiment140 dataset
- Overall brand sentiment for both brands and its change over time
- Data Visualization
- Project Presentation
- Any source code, file, and resource used for the project

6. REFERENCES

[1] snscrape, Sep 2020.

[2] Tweepy api, Sep 2020.

[3] F. Bravo-Marquez, M. Mendoza, and B. Poblete. Combining strengths, emotions and polarities for boosting twitter sentiment analysis. WISDOM '13, New York, NY, USA, 2013. Association for Computing Machinery.

[4] A. Go. Sentiment classification using distant supervision. 2009.

[5] S. Guest. Sap brandvoice: Ps4 vs. xbox one: Winner emerges via social media analysis, Nov 2013.

[6] KazAnova. Sentiment140 dataset with 1.6 million tweets, Sep 2017.

[7] A. Mittal and S. Patidar. Sentiment analysis on twitter data: A survey. In *Proceedings of the 2019 7th International Conference on Computer and Communications Management*, ICCCM 2019, page 91–95, New York, NY, USA, 2019. Association for Computing Machinery.

[8] S. M. Mohammad, P. Sobhani, and S. Kiritchenko. Stance and sentiment in tweets. *Special Section of the ACM Transactions on Internet Technology on Argumentation in Social Media*, 17(3), 2017.