LIN 127 Project Report

Munir Sayani
mmsayani@ucdavis.edu

Sam Shahriary
sshahriary@udavis.edu

Muhammad Khan
mshkhan@ucdavis.ed

Abstract

The purpose of the project is to use NTLK analysis to learn which topic is most popular on Twitter on a normal day. The categories chosen were News/Politics, Sports, Music, Entertainment, and Science & Technology since we believe that these encompass almost all the topics people tweet about. The twitter API was used to gather tweets from various twitter accounts that primarily tweeted about a particular category. The twitter accounts chosen for these categories were determined by carefully analyzing many tweets within those accounts and then deciding if they were a good fit for a category. Afterwards, we gathered a corpus from the twitter live feed and determined which categories are most tweeted about on a daily basis. The findings stated that News/Politics was talked about the most and entertainment was the least.

Primary authors: Khan

1 Credits

Our three man group is comprised of all Computer Science majors so our skills are best fit for the coding portion of this project. Sam Shahriary started us off by coding a script to gather data from the Twitter API. Sam then used his Linguistics knowledge to create training, development and a testing set.

All three of us made a list of twitter accounts which would be a good representative of a category. We then participated in many discussions to make sure that the chosen accounts would be a good source of our training and development sets. Apart from that, we also discussed how many tweets should be used in both sets as well.

Muhammad Khan and Munir Sayani worked together to develop scripts to create a model from the training set. Then used the model to classify the development and testing set.

Primary author(s): Sayani, Khan and Shahriary

2 Introduction

As we started researching for this project, we ran across a list of the most followed twitter accounts of all time. We realized that these accounts are mostly comprised musicians, actors, politicians and other celebrities. It was impressive seeing this diversity in occupation of these accounts which led us to another query. It peaked our interest to see what topic was most tweeted about? Initially this query felt too narrow. We broadened it by focusing on the subject of a tweet; Is it focused on music, politics, news or something unexpected?

We started our discussion by attempting to generalize categories so that we could categorize each tweet by topic. The categories we decided upon were News/Politics, Sports, Music, Entertainment, Science & Technology. We did predict some issues with Entertainment since it was a too broad. However, we felt our assessment would still be accurate since our other categories were still narrow.

Primary author(s): Sayani

3 Tasks

We divided our project into multiple tasks. The first task is gathering the data from Twitter and creating a training set. We then create a model using NLTK which we use to classify our tweets into categories. Afterwards, we create a diversified development set to test accuracy, precision, recall and F-score. Lastly, we gather a testing set to determine the percentage of tweets for each category. The testing set would then encompass live tweets throughout the day. This will allow us to gather which category is most popular on twitter.

Primary author(s): Sayani

4 Data

The Twitter API provided much flexibility including plenty of tweets to request in a given time period, instructions on how to authenticate the API, and how to use API functions given Twitter's complex platform. The main function used was the GET statuses/user_timeline function. This function, given an account's username and set of preferred parameters, returned a list of status/tweet objects that held various metadata about a tweet. The tweets were pulled from users who best fit our categories.

When building our training data we chose accounts that we thought would best fit our categories to pull from. For example, 'News' pulled from accounts such as 'DailyMail' (national) to 'RT_com' (international), and 'Sports' pulled from accounts including 'SportsCenter', 'Olympics', and many other accounts representing different sports. We tried to maintain a diverse and accurate group of accounts to pull tweets from. We wanted our training set to be large and broad, therefore we pulled a total of 12,000 tweets from various accounts in each category.

The development and test set used the same method but were pulled from different accounts. To ensure the fairness and accuracy of our model, we wanted to keep testing on different accounts that spanned the spectrum on these broad categories.

After going through numerous twitter accounts and tweets I noticed a lot of limitations and problems that could arise in the building of our classifier model. One limitation was that most of our tweets that were grabbed from each account were X amount of most recent tweets posted. This means that our classifier would be mainly built off of words used in events that occured in a specific range of time (there might be more sports tweets about basketball if a big game had just occurred, scarce amount of tweets about summer Olympics in between 4 years it's not occuring, etc.) and would be inaccurate if tested on tweets during a different time. Solution would be to build training set over years and update classifier based on new sports, news, scientific breakthroughs. A second limitation was the overlap of categories (as mentioned below). Many news sources on twitter incorporate politics and science/tech. Entertainment usually incorporates music. I think to solve this issue we'd need more defined lines between our categories that better fit reality.

The last limitation I noticed was that tweets were becoming more visual than textual. The attention span of people is getting shorter with social media which results in more attention grabbers by tweets. I saw many sports, news, entertainment, and even scientific tweets that had a few foreshadowing or clickbait headlines that directed users to a video rather than having a textual explanation. This could result in our classifier grouping all tweets into wrong categories given that they're only represented by clickbait titles.

Primary author: Shahriary

5 Results

After going through our development set, we got an accuracy of 58%. Since accuracy alone is not a good measure of our classifier, we went ahead and calculated the precision, recall and F-score for each of our category. The results are listed in the table below.

Category	Precision	Recall	F-score
News/Politics	74%	78%	0.76
Sports	61%	64%	0.62
Music	62%	63%	0.62
Entertainment	63%	46%	0.53
Science & Technology	58%	67%	0.62

Table 1. Development Set Data Results

Primary author: Khan

After going through our test set we got the following results:

Category	Percentage
News/Politics	24%
Sports	16%
Music	22%
Entertainment	16%
Science & Technology	22%

Table 2: Test Set Results

6 Analysis

We expected news to be have the highest precision and recall because it has the most amount of words used mainly in tweets regarded as news and politics related. Words like ‘president’, ‘congress’, ‘parliament’ or even country names are used primarily when talking about news and politics. Also, many world leaders like President Trump and Putin are known and mentioned by people and we expected those names as a classifying factor as well.

Similarly, for Science and technology, there are so many unique words that ensure a higher probability. Words like ‘space’, ‘satellites’ and even names of various gadgets and technology aids a lot in classifying something in this category.

The poorest precision and recall scores occur when our program tries to classify tweets that are involved with entertainment. Our entertainment category consists of movies, TV shows and celebrity gossip (but not

something as drastic as celebrity deaths since that is News). We believe that this category has a lot of intersection with music since all the big artists can be considered as celebrities as well which probably skewed up our results. Also, we made our training set during a time when 'iHeart Music Awards' were occuring which probably made music related topics one of the highest trending topics in general, and therefore, the twitter accounts that solely focus on celebrity life tilted more towards talking about musicians.

However, with sports, we expected better results since there are a lot key words that would be used primarily to define it. Those words include sports specific terms like 'goal' or 'jab' as well as the names of various sports as well as the famous athletes themselves. We believe that the results are not that good because sports can be a form of entertainment. Twitter accounts that go over celebrity lifestyles and gossip go over sports athletes as well since they are some of the most followed types of athletes.

Despite all the reasons that could have skewed our results, our results show that News and music are two of the most tweeted about topics on twitter. Twitter is a good means for sharing news and on the other hand people in general talk about music, especially during a time when music awards are taking place.

Primary author: Khan

7 Conclusion

Overall, we feel that the project was successful as we were able to determine which topic was most talked about on Twitter (News). We were a bit disappointed that the accuracy was below 70%. After our results, we discussed possible solutions to extend this work and receive better results.

First, our "Entertainment" category was too broad. As mentioned earlier, Entertainment overlaps Sports and Music. Thus, a better way to categorize is to eliminate Entertainment and replace it with subcategories like Movies, TV Shows and Celebrities. Doing this will help to better focus our categories on a singular theme and avoid overlap. We can gather corpora of these categories and ultimately achieve a more accurate result.

Also, we can focus on extending our training set. Although, we have over 12,000 files in the set. It would be better to increase the corpora. Furthermore, we can diversify the training set with more unique topics, such as Gaming.

Lastly, we can gather our Twitter set from a larger time frame. As mentioned earlier we spent one day gathering data for our training set. This forced our data to be heavily biased toward whichever topic was trending, which in our case it was 'iHeart Music Awards'. This negatively affected our corpora by confusing Entertainment and Music. Therefore, the best way to avoid this issue is to develop a training set over a longer period of time.

Primary author(s): Sayani

Reference

https://twitter.com/?lang=en

https://developer.twitter.com/en/docs/tweets/timelines/api-reference/get-statuses-user_timeline.html

https://developer.twitter.com/en/docs/tweets/timelines/guides/working-with-timelines

Google Drive Link: https://drive.google.com/file/d/1OJf5o2Y-LnAtiZXOA4T9hlwXOG9Wc59X/view?usp=sharing