CSC 555 (601) : TERM PAPER

Team:
Prashant Mohan Srivastava (pmsrivas)
Ruchika Verma (rverma5)

***Claim:* Identifying correlations between the trends in social media users' sentiments and real world events. Does this correlation exist?**

What do we mean by sentiment analysis?

The process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc., is positive, negative, or neutral.

Humans are known for using their freedom of speech and what other better way to do so than commenting on public forums and social networking sites! Thus, any social networking site today provides an immense amount of data from which significant information can be extracted. One such platform is YouTube, a video sharing website. In addition to sharing of videos, YouTube provides features to subscribe, upvote, downvote and comments on videos. Comments on YouTube are a potential source of data that can be put to analysis in numerous ideas and remain the focus of this project.

This paper elaborates our process and technique for the implementation of the idea and our attempt to verify the claim. The project is primarily divided into 3 parts: Data Collection, Sentiment Polarity Calculation, and Correlation Analysis. We take up each part and discuss the technique employed and the observations made.

Data Collection: Data from social network or social media forms the foundation of any sentiment analysis project. In our project, the basic dataset was the comments on YouTube. With the help of Google Developer Tools which provided the APIs to access YouTube's data, we were able to collect about 15,000 comments for out three search phrases. We decided to collect and analyse the data on the three most popular tennis players, Roger Federer, Rafael Nadal and Novak Djokovic. The search phrases we used for getting the list of videos for each of these players were "Roger Federer over the years", "Rafael Nadal over the years" and "Novak Djokovic over the years". These search phrases allowed us to fetch the videos on each player spread over several years.

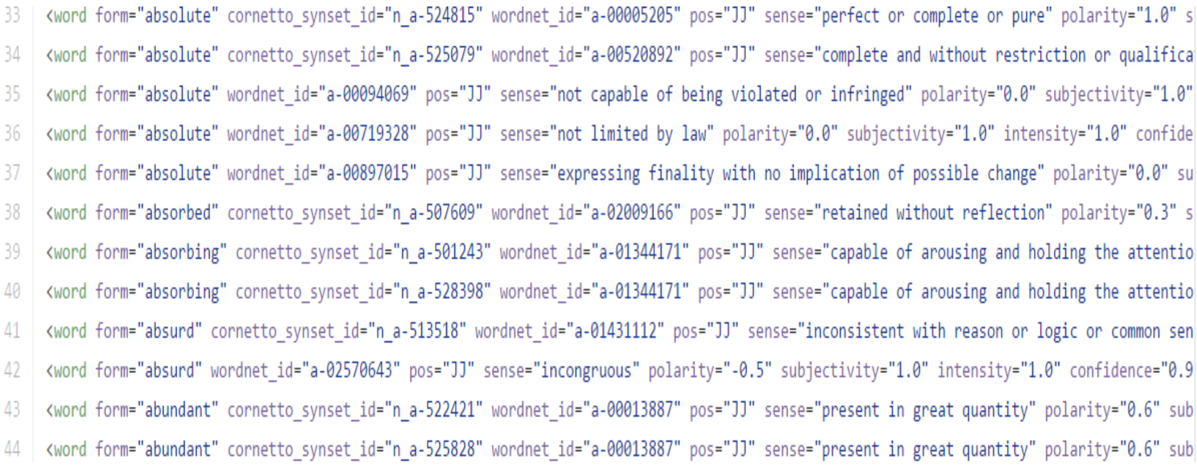
Next, we extracted the top-comments and their timestamps for each of the videos on each of the three search-phrases. All the comments for a video were stored in a file in JSON format which efficiently captured all the data that would be required.

This phase is one of the most important part of the project and therefore the challenges we faced were also difficult. As YouTube has restricted its APIs for developers, we could only fetch 50 videos for each search phrase and 100 comments for each of those videos. Since we know that majority of the comments on YouTube are hardly of much use, we had to extract the ones we thought we could use for the project. Therefore, this limitation posed a serious roadblock to the analysis.

Sentiment Polarity Calculation: Once we had the dataset that we could use for the analysis, the next task was to calculate the sentiment polarity for each of these comments. A standard tool for doing this is a Python package called NLTK (Natural Language Toolkit). NLTK provides almost all the essential tools for natural language processing and is widely used for sentiment analysis. Therefore, we chose to use NLTK for our project.

Since getting NLTK to work on a dataset can be quite challenging, we decided to make use of TextBlob, a package in Python which is a wrapper on the functionalities of NLTK. TextBlob has its own corpus which can be directly used for calculating sentiments of statements. TextBlob assigns a polarity rating to a statement ranging from -1.0 to +1.0 where -1.0 is the most negative and +1.0 is the most positive. For example, in the image above the word absolute when used in different “senses” as defined below in Fig 1 has different polarity values (0.0 & 1.0). It handles the cases of negation and modifier word also. For instance, the word “great” is given a polarity of 0.8 while the words “not great” is given the polarity of -0.4 and the words “very great” is given a polarity of 1.0.

Once we have calculated the polarity of each comment on all the videos from a search phrase, we create a csv file with the dates and the polarities of all the comments.



```
<word form="absolute" cornetto_synset_id="n_a-524815" wordnet_id="a-00005205" pos="JJ" sense="perfect or complete or pure" polarity="1.0" s
<word form="absolute" cornetto_synset_id="n_a-525079" wordnet_id="a-00520892" pos="JJ" sense="complete and without restriction or qualifica
<word form="absolute" wordnet_id="a-00094069" pos="JJ" sense="not capable of being violated or infringed" polarity="0.0" subjectivity="1.0"
<word form="absolute" wordnet_id="a-00719328" pos="JJ" sense="not limited by law" polarity="0.0" subjectivity="1.0" intensity="1.0" confide
<word form="absolute" wordnet_id="a-00897015" pos="JJ" sense="expressing finality with no implication of possible change" polarity="0.0" su
<word form="absorbed" cornetto_synset_id="n_a-507609" wordnet_id="a-02009166" pos="JJ" sense="retained without reflection" polarity="0.3" s
<word form="absorbing" cornetto_synset_id="n_a-501243" wordnet_id="a-01344171" pos="JJ" sense="capable of arousing and holding the attentio
<word form="absorbing" cornetto_synset_id="n_a-528398" wordnet_id="a-01344171" pos="JJ" sense="capable of arousing and holding the attentio
<word form="absurd" cornetto_synset_id="n_a-513518" wordnet_id="a-01431112" pos="JJ" sense="inconsistent with reason or logic or common sen
<word form="absurd" wordnet_id="a-02570643" pos="JJ" sense="incongruous" polarity="-0.5" subjectivity="1.0" intensity="1.0" confidence="0.9
<word form="abundant" cornetto_synset_id="n_a-522421" wordnet_id="a-00013887" pos="JJ" sense="present in great quantity" polarity="0.6" sub
<word form="abundant" cornetto_synset_id="n_a-525828" wordnet_id="a-00013887" pos="JJ" sense="present in great quantity" polarity="0.6" sub
```

Fig 1: TextBlob’s sentiment.xml file from github.

Correlation Analysis: After the polarities of all the comments were obtained, we grouped all the polarity values for the comments which were posted on the same day. We used pandas package in Python for this task. The data was grouped by the date and median of polarity values was taken in order to find the central polarity value for each day. Next we plotted the graph of this data with the dates vs polarity values.

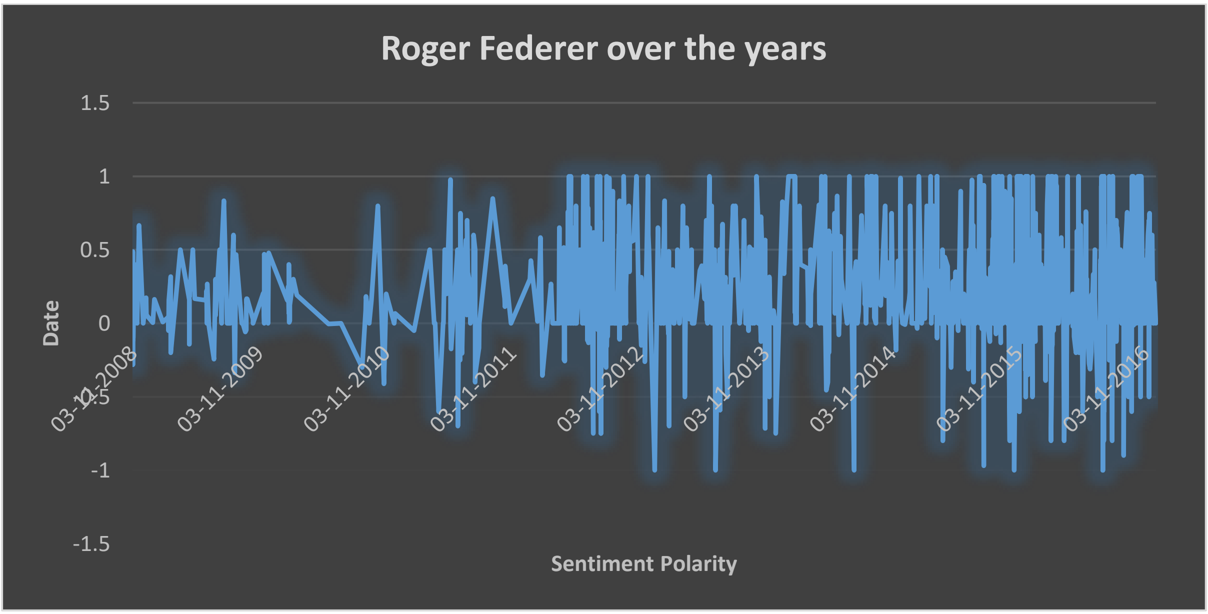


Fig. 2. Sentiment Polarity trend for Roger Federer.

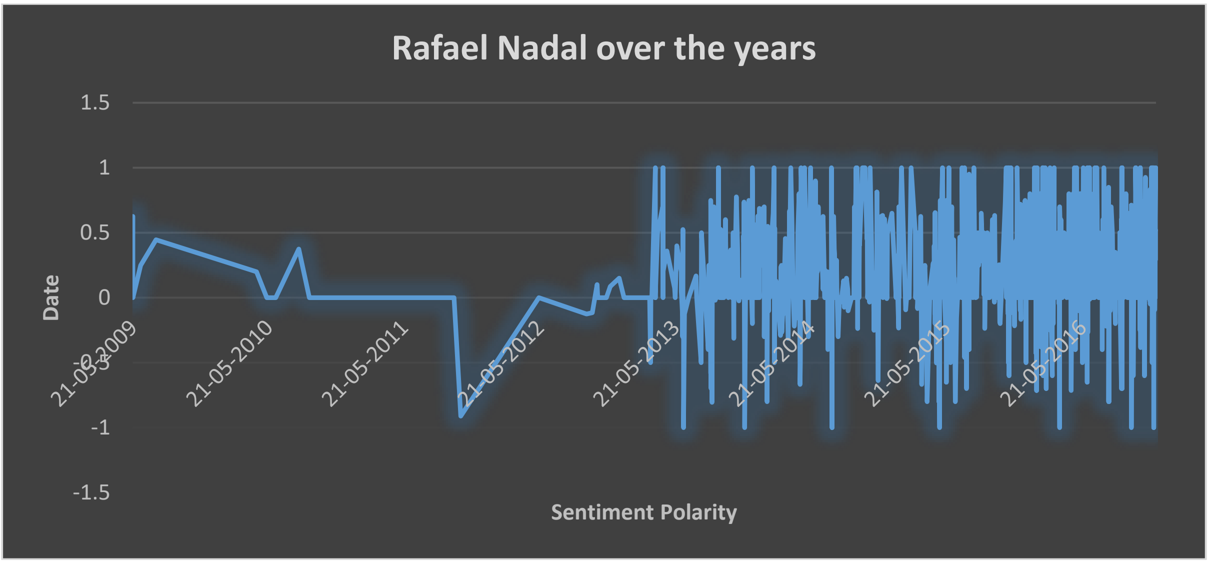


Fig. 3. Sentiment Polarity trend for Rafael Nadal.

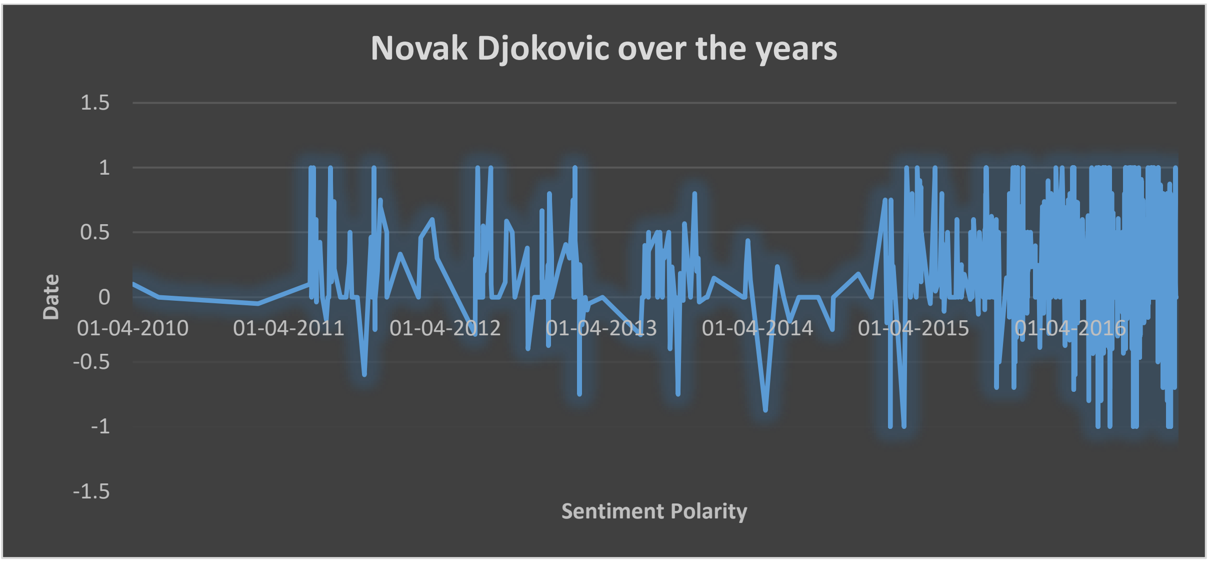


Fig. 4. Sentiment Polarity trend for Novak Djokovic.

From the graph plots shown above, we observed few instances and sections where the polarity trend showed some rise or decline and we tried to find an event in the players' career which may be an indication for such trends. For finding such events we trusted Wikipedia and other news on the internet.

Observations: Following are the few points that we observed from the above graphs:

1. We notice that all the graphs are sparse on the left side and highly dense on the right side. This may be attributed to the fact that although these players made their entry into tennis at different point of times, they have attained a legendary status in sports in the last few years. Therefore, their videos on YouTube now attract more comments than before.
2. In Wimbledon 2015, Novak Djokovic won the title and Roger Federer finished as runner-up. When we observe fig 2 and fig 4, we notice that during the June-July period of 2015, there is a decline in the positive polarity of Federer, whereas for the same period the polarity for Djokovic is mostly positive.

3. In fig 3, we observe a sharp decrease in the polarity trend of Rafael Nadal from 2011 to 2012. This can be related to the fact that Nadal was out of the game for a significant part of the year due to injury.
4. Rafael Nadal beat Roger Federer in the Australian Open and Wimbledon of 2009. We can observe a decline in the graph of Federer for the time after 2009 till 2010 and a slight increase in the graph of Nadal after 2009.

Limitations: While working on this project, we encountered following limitations:

1. The problem statement aims at finding if the correlation exists between the real world events and the users' sentiments on YouTube. Since the sentiments can be expressed in multiple ways and on multiple subjects, it could be inconclusive to tell if the correlation exists. This finding is significantly dependent on the researcher and his/her observations.
2. Our methods of data collection are entirely done through YouTube's APIs. Currently Google allows only 50 videos for each search phrase and only 100 comments for each of these videos. With these limitations, we could easily lose significant amount of relevant data. Since, the amount and quality of data has huge impact on the findings, we cannot be fully sure of the results.
3. The findings of this projects relies heavily on the data collected as well as the approach used for the sentiment analysis. Since this analysis is totally done by predefined algorithms, the accuracy of the results is very much dependent on these algorithms. Algorithms such as Naïve-Bayes and Support Vector Machine offers good solution to this problem, but they are still not very high on accuracy.

Prerequisites for applying these finding to practice: The results of this project are derived from observations which were made while constantly keeping the problem statement in mind. Therefore, with the given data, what is it that we must look for is a crucial step. The exact problem formulation is the most basic and essential prerequisite. The decision of the period of time which has to be taken into observation should be decided prior to collection of the data so that relevant data can be retrieved from the APIs.

Conclusion: As we know that social media and social networks are open platforms where people from vast and varied demography post and share ideas. Since there is little control over how these ideas are shared, we often tend to find a mix of opinions which are sometimes relevant and sometimes not. For instance, in the data collection phase of or project, we came across innumerable comments which were not at all related to the player. Getting rid of such comments was a difficult task. From this we learnt that not all the opinions from social platforms can be used for analysis. In the analysis phase, we came across several inconsistencies in the polarity trends. Although we observed slight trend in some parts of the graph, in other parts the data was mostly inconclusive. With our observations on this project, we can conclude that the opinions expressed on social media possess significant potential for extracting useful information and also serves as useful resource for interesting studies, but these opinions cannot be fully trusted as they lack structure and sometimes relevance. Also, in order to extract full potential of these opinions, better algorithms for natural language processing with higher accuracy.

References:

1. Polarity Trend Analysis of Public Sentiment on YouTube: http://home.engineering.iastate.edu/~zambreno/pdf/KriZam13A.pdf
2. Link for the whole sentiment.xml file on github: https://github.com/sloria/TextBlob/blob/eb08c120d364e908646731d60b4e4c6c1712ff63/textblob/en/en-sentiment.xml
3. Roger Federer career statistics: https://en.wikipedia.org/wiki/Roger_Federer_career_statistics
4. Rafael Nadal career statistics: https://en.wikipedia.org/wiki/Rafael_Nadal_career_statistics
5. Novak Djokovic career statistics: https://en.wikipedia.org/wiki/Novak_Djokovic_career_statistics