Foundations of Data Science Final Project
Overwatch League Twitter Analysis

Jiawei Wang
University of Alabama at Birmingham
jwang96@uab.edu

Shuhui Wu
University of Alabama at Birmingham
shuhui@uab.edu

Feng Zhang
University of Alabama at Birmingham
feng1013@uab.edu

ABSTRACT

In this project we have crawled two months of Overwatch League related data (67.6 MBs of text) from Twitter, using methods like *wordcloud*, *sentiment analysis* and *language-based analysis*, trying to extract the characteristic of the leagues and how people support their beloved teams.

1 INTRODUCTION

Overwatch League, also called as *OWL*, has recently opened its 2019 Seasons which has a prize pool of 1.5 Million dollars for the winning teams. We decided to analyze each team's performance by evaluating user's tweets about the related topics.

2 DATASET

Since Twitter Standard API does not support querying tweet one weeks ago, we use PyQuery to manually crawl our data via Twitter's web-based search API. Our data ranged from February 14, 2019 (when the current season started) to April 21, 2019 (when the project begun), which contains 393780 tweets from all 20 team's name as search queries. Here is a snippet of one of the tweet data:

```
{'id': '1119752509789097984',
 'date': 'Sat Apr 20 18:59:36 +0000 2019',
 'username': 'BQB_kr',
 'retweets': 0,
 'favorites': 17,
 'text': 'ggs @ShanghaiDragons',
 'language': 'und',
 'mentions': '@ShanghaiDragons',
 'hashtags': ''}
```

3 ANALYSIS

3.1 Wordcloud

After crawling tweets for each team by their names, it is important to check if the data has been crawled correctly since some of the names might be as well a popular word used commonly or they are not how fans usually called them. We can create word clouds for each of the crawled data, and analyze if anything goes wrong.

3.1.1 *"Open Sesame".*

Atlanta Reign. As we can see from the wordcloud (Figure 1), the largest word is "dafran", a member of *Atlanta Reign* who carries most of its match. The second frequent word is 'LetItReign', the slogan of the team. We might say the crawl data represents well for the team.

Figure 1: Wordcloud: ATLReign

Figure 2: Wordcloud: NYXL

Figure 3: Wordcloud: Shanghai Dragons

NYXL. "NYXL" is an abbriviate for the team *New York Excelsior* (see Figure 2), which almost exclusively represents the team. In fact, this is a strong team that hardly ever losing a game (although it has just lost its first game recently). No wonder the word "win" is the most frequent word in the tweets.

Shanghai Dragons. Like the *opposite* of NYXL, *Shanghai Dragons* is considered as a weak team since it had never won before (actually, several members and staffs were departed the team after the last season, since they had literally no wins in the whole season 2018). In this season they had changed nearly all its members and

Figure 4: Wordcloud: Houston Outlaws

Figure 5: Wordcloud: Philadelphia Fusion

Figure 6: Wordcloud: WashJustice

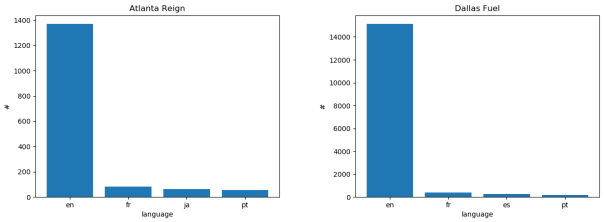
had its first win shortly after the season begun. Hence words like "Breakthrough", "first win" have reasonably appeared on the word cloud (Figure 3).

3.1.2 "Sorry, wrong one".

Huston Outlaws. It seems that the fans have another name for *Huston Outlaws*, but definitely not "Huston Outlaws". From the word cloud (Figure 4) we can see most of the frequent word is just other teams' names, which because most of the tweets are game reports from news accounts. When we manually looking for the *correct callsign* for the team, we found that the word "Outlaws", which is the correct one how fans called them, is actually abused by a recently released game from *DC*. That means we cannot use the correct word to crawl the data since the result will be *heavily polluted.*

Philadelphia Fusion. "Fusion" is used widely by *Philadelphia Fusion* fans, and it is also the official account name for the team. However the word is already a *common english word* and frequently used worldwide. We desparately found that the only one who likes calling the full name of the team is just a generous sponsor who likes *giveaway gaming chairs* when the team wins (Figure 5).

3.1.3 "Eureka!"



(a) Atlanta Reign **(b) Dallas Fuel**

Figure 7: Language distribution for both teams

Washington Justice. Washington Justice is the first team we managed to correct when we investigating word clouds. Before we used the correct keyword, "Washington Justice" will be automatically corrected to *Washington Post* and *Justice Department* by Twitter correction mechanism. Hence results like *the Muller* expells the OWL content and shows on the word cloud (Figure 6). Since *daily news* has a more proportion of the population compared with Overwatch, we can hardly see any result from the team in the word cloud. Hence we investigate tweets from other teams, figured out that user tends to use "WashJustice" to imply the teams, we changed our search query to that and finally got the corrected result.

3.2 Team-Language Distribution

The language distribution is intended to find the regional composition of fans of each team since we cannot acquire the exact location information from Twitter which obviously is the violence of the user privacy policy. Therefore, we compromise that we use the language type instead to represent the regional difference between users.

We go through the dataset we get from crawler API where each tweet datum is organized in JSON format which is convenient to be converted to python *dict* class. We filtered the content in "language" and maintain another *dict* to store the count of different languages in tweets related to different teams. We plot the data in 2 format, different language distribution regarding a team and percentage of each team regarding a language.

3.2.1 Language distribution in Teams. Unfortunately, we cannot find any differential in team view. It turns out that English is the most dominant language, however, other languages don't have any special features (see Fugure 7).

3.2.2 Team distribution in Languages. However, in the language distribution view, we can find some interesting features that some of which we can think of the reason to explain and some are not. In the team distribution of French (Figure 8), we can see that the most French-speaking supporters are the fans of *Paris Eternal* from which we can assume that normally fans tend to support the local team. Another proof is the *Seoul Dynasty*, in which Korean fans have shown their stronger anticipation and the heat of discussion on Twitter (Figure 9).

3.2.3 Heat team analysis. Another discovery is that no matter where people come from, people tend to support the *strong team* like *NYXL* not only because their advance in-game skill, but the

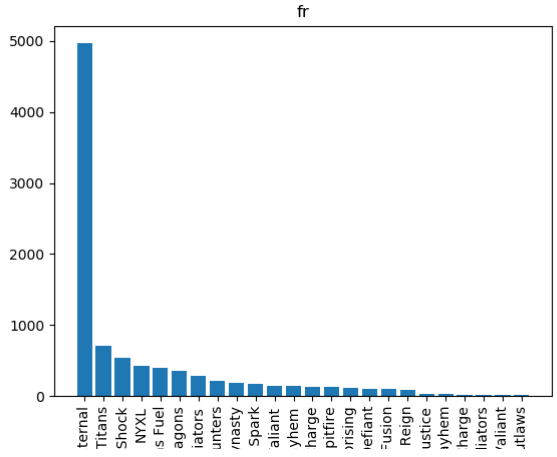


Figure 8: Team distribution for French

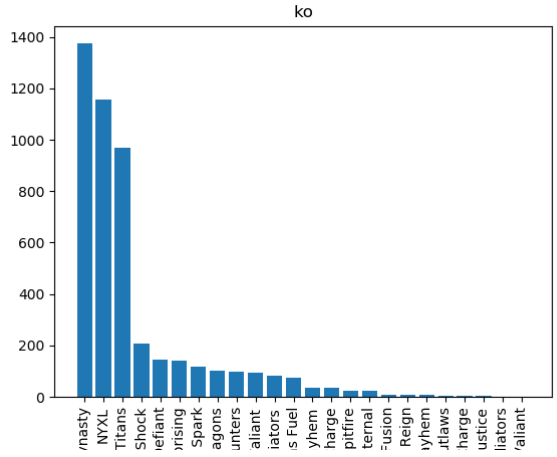


Figure 9: Team distribution for Korean

attention they have and the influence they have on the normal players that don't usually watch the games, which is another type of fashion generated from e-sports industry that shares common features to traditional sports.

The heat of discussion regarding a team is not only related to the power of the team. For example, in English chart (Figure 10), the most heated team other than NYXL is *Shanghai Dragon* which is hard to explain the reason why but under a special circumstance that *Shanghai Dragon has achieved their first win* since it was founded. It was a rare situation that Overwatch League has been founded for two years, however, there is a team that has never won any game of it for that long.

3.3 Sentiment Analysis

3.3.1 Methodology. *TextBlob* is a Python library for processing textual data. It provides a consistent API for diving into common

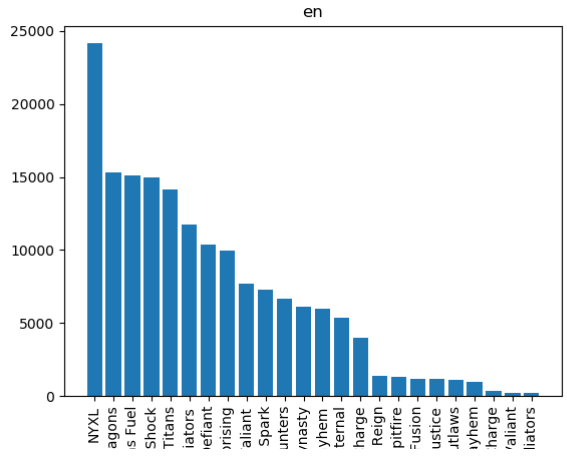


Figure 10: Team distribution for English, where NYXL as #1, and Shanghai Dragons as #2.

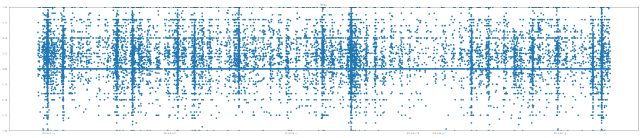


Figure 11: Sentiment: NYXL

natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, and more.

We used its sentiment analysis functions. Take a string as input and get a value $[-1, 1]$. If a value greater than 0 means this text is positive, otherwise, it is negative.

We extract "text" and "date" from JSON files, which are the twitters data we got. And analyze text with TextBlob. In this way, we get a lot of (date, polarity) pairs. Then, draw a scatter plot with this data for each team.

3.3.2 Analysis. We got 20 scatter plots for 20 teams. Although their data volume is not the same because of many reasons, their distribution is very similar. Some teams have more data because they are very popular, such as *New York Excelsior*. However, some teams are also very popular but they only have a little data in our project, such as Houston Outlaws and Philadelphia Fusion. There are some typical plots below.

In addition, till now Season 2019 is still going, therefore the analysis only covers Stage 1, which begins from February 14 to March 17, and Playoff games from March 21 to March 24, and parts from Stage 2, which begins from April 4 to April 21, where the crawling stops.

NYXL. *NYXL* has numerous fans because this is one of the strongest team and they are an old team who attend OWL last year. Meanwhile, it's a U.S city team, which make them own lots of local fans in the U.S. They earned most matches in last season. All these reasons make them extremely popular all the time. It's obvious that most twitter are positive. (Figure 11)

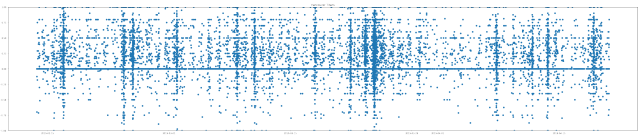


Figure 12: Sentiment: Vancouver Titans

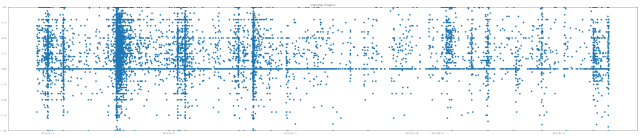


Figure 13: Sentiment: Shanghai Dragons

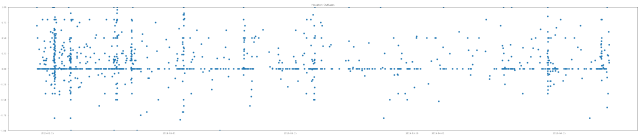


Figure 14: Sentiment: Houston Outlaws

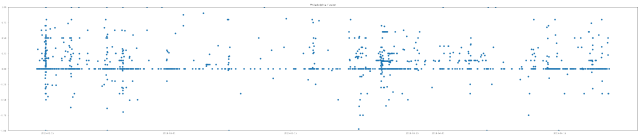


Figure 15: Sentiment: Philadelphia Fusion

Vancouver Titans. *Vancouver Titans* is another candidate for 2019 season champion. They are the team who won most matches so far. Comparing to NYXL, they have fewer data. Maybe because they are the new team who attend OWL 2019. In the plot, there is a peak period around 2019-3-24. This is the Stage 1 playoffs time and Titans is the champion. The twitter related to Titans are almost all positive. (Figure 12)

Shanghai Dragons. As an old team with no win, *Shanghai Dragons* captured much attention. They lost 40 matches last season and finally got the first win in Stage 1 in 2019. The peak period in the plot below is around 2019-3-22. This is a milestone for Dragon and their fans. It means huge in OWL so that there are such amount of twitters are talking about Dragon. Though the data in peak period have many positive twitters, there are still many negative data. Dragon is controversial obviously. (Figure 13)

Dragons has the only female player in OWL so that they earned a lot of female fans. However, Dragon is a team in a Chinese city, their Chinese fans cannot use Twitter in the mainland. Thus, the data concerned about Dragon is less.

Houston Outlaws & Philadelphia Fusion. Actually, these two teams have countless fans. As the only U.S team whose players are all from the U.S, *Outlaws* is almost the team who has most American fans. *Fusion* is very strong and they would almost be the champion by last season. However, by the date we obtained our data, these two teams are very abnormal. (Figure 14, Figure 15) After consideration, we found this may because of the ambiguous keyword. For example, there is a new game named “outlaws” will be released.

4 CONCLUSIONS

In *Wordcloud*, we successfully using the extracted information to justify the correctness of our data. Also, we can easily characterize each team by the words showing on the word cloud.

In the *language distribution*, we have revealed that although heat teams would gain a significant proportion of attention, people would also tend to support their local teams.

In terms of Sentiment analysis, The distribution is very similar no matter how many data each team has. Positive twitters are much more than negative twitters. The distribution is associated with the date because of the match schedule: it’s not continuous rather intermittent. The peak period of teams is related to their win and lose.