# CASE STUDY 1: COLLECTING DATA FROM TWITTER

**TEAM 12**

## DS 501: INTRODUCTION TO DATA SCIENCE


**TEAM MEMBERS**

Chu Wang

Saranya Manoharan

Rishitha Kiran

Di You

Valerie Tuzel

# 1. MOTIVATION

The Irish airline Ryanair recently announced that they were going to cancel up to 50 flights a day until the end of October, because their punctuality went down from nine out of ten flights to eight out of ten flights. The reason given was “the shortage of staff”. Ryanair's marketing director, Kenny Jacobs, said: "We have messed up in the planning of pilot holidays and we're working hard to fix that." The “tighter crewing numbers”, said to be a result of a new leave calendar, means pilots and cabin crew need urgently to take vacations [1]. On each of Saturday, September 16 and Sunday, September 17 more than 80 Ryanair flights were cancelled. As a result, on each of these days, as many as 15,000 Ryanair passengers were told shortly before their flights departure, that their plane had been cancelled. They were then asked to choose between claiming a refund or rebooking their flight [1].

Because of this reason a lot of people were tweeting about their suddenly cancelled trips during the weekend of September 16 and their frustrations regarding the same. We utilized this opportunity and performed data analysis on the Ryanair data collected from Twitter using the Streaming API. All the tweets related to this were collected during that time was preprocessed and cleaned. After data cleaning, the dataset was analyzed accordingly and results were tabulated. Additionally, with the obtained dataset sentimental analysis was performed in order to identify and categorize user opinions.

# 2. DATA COLLECTION

## 2.1. Sampling Twitter Data with Streaming API about a certain topic

We used Twitter’s Streaming API and for every 1000 tweets data was saved to a JSON file and the JSON files were then combined as one dataset. The process of pulling data was constrained with time limit because of the API rate limits. Moreover, the tweets were collected based on keyword “ryanair” and overall, we collected 6,000 tweets.

# 3. DATA ANALYSIS

The following section will discuss in detail regarding the different analysis that were performed over the Ryanair dataset.

## 3.1. ANALYZING TWEETS AND TWEET ENTITIES WITH FREQUENCY ANALYSIS

For our data analysis, we made use of NLTK (Natural Language Tool Kit) in order to process the tweets. The results of the experiments that we ran on the data we collected are shown here.

### 3.1.1. Word Count

Using the tweets collected in the previous step, we computed the frequencies of the words that were used in the tweets that we collected.

The following is the step by step procedure of the word count process:

- We extracted all the text entity in the collected tweet data and filtered the words that were in English language.
- From the obtained tweet text, we performed splitting of valid English words such as “cancel”, “is”, “it”, and, “leave”.

- The word list obtained in the previous step contained lot of stopwords such as “it”, “is”, and, “not”. Therefore, we used NLTK library to make a list of stopwords.
- Then, using NLTK we filtered out all stopwords and performed additional filtering to remove URL’s, punctuations, numbers, hashtags and user_mentions from the list of words computed earlier.
- Next, we performed word count on the resulted word list of the previous step and determined their frequencies.
- Finally, we made a table of the top 30 words and their frequencies as shown in Table 1.

Top 30 Words With Their Counts:

| Word | Count |
|---|---|
| cancelled | 736 |
| list | 337 |
| cancels | 313 |
| next | 302 |
| get | 257 |
| day | 256 |
| pilots | 219 |
| cancellations | 216 |
| know | 215 |
| rights | 185 |
| leave | 183 |
| cancelling | 180 |
| people | 173 |
| dont | 171 |
| messing | 170 |
| days | 167 |
| weeks | 166 |
| entitled | 159 |
| full | 158 |
| losing | 156 |
| us | 150 |
| brexit | 149 |
| forget | 147 |
| workers | 146 |
| annual | 141 |
| customers | 138 |
| talk | 138 |
| going | 127 |
| please | 121 |
| one | 119 |

**Table 1: Table of top 30 words with their counts**

### 3.1.2. Find the most popular tweets in your collection of tweets.

The next problem was to acquire the most popular tweets including the word “ryanair”. The most popular tweets are those tweets which are the most re-tweeted tweets and which have the maximum retweet count. The following are the step by step procedure followed to determine the top 10 popular tweets:

- From the tweet dataset, we first filtered the tweets that contained “retweeted status” entity and “retweeted count” greater than zero.
- Then, we retrieved the tweet text from the text entity along with their count and stored in a list.
- Then, we sorted this list of retweets in descending order.
- Finally, we extracted the top 10 tweets from the sorted list and made a table of the tweets and their retweet count as shown in Table 2.

Top 10 Tweets:

| Count | Text |
|---|---|
| 1062 | RT @Ryanair: For a chance to WIN a €/£100 voucher simply FOLLOW, RETWEET & tell us which destination you stopped on below using… |
| 434 | RT @Channel4Racing: VAUTOUR with @Ruby_Walsh on board finds an extra gear to win the Ryanair Chase! #CheltenhamFestival https://t.co/ezSqhc… |
| 136 | RT @LeaveEUOfficial: When you talk about losing workers' rights after Brexit but forget your pilots are entitled to annual leave... 🙄… |
| 135 | RT @PaulWoolford: How is it legal that @Ryanair can cancel 400,000 seats on it's flights yet there is no way to cancel a Ryanair seat & get… |
| 124 | RT @MrHarryCole: While lecturing about politics. https://t.co/OK5fXVIFZ5 |
| 115 | RT @_chloemo: @Ryanair This isn't good enough. Release a list now of all the flights you plan to cancel so your paying customers… |
| 111 | RT @SimonCalder: Ryanair cancellations: up to 400,000 passengers affected. Your rights here. Let me know if you have other questions. https… |
| 109 | RT @ollieoioioi: If your flight is cancelled by #ryanair DO NOT accept the refund. You can sue them for compensation for not giving 2 weeks… |
| 108 | RT @BBCWorld: Ryanair cancels flights after 'messing up' pilot holidays https://t.co/kfY5iUCBPq |
| 104 | RT @patrick506: @Ryanair Publish a full list now of all the flights you intend to cancel, save yourselves some goodwill, possibly. @Ryanair |

**Table 2: Table of top 10 most popular tweets**

### 3.1.3. Find the top 10 hashtags, top 10 user mentions that are the most popular in our collection of tweets

In this section, we determined the top 10 hashtags and the top 10 user mentions. The top hashtags are those hashtags that have been used the most. Similarly, the top 10 user mentions are the 10 most mentioned users, in our collection of tweets. The following is the step by step procedure of the process:

- From the collection of tweets, we extracted all the hashtags and user_mentions and stored them in a list along with a count of how many times they appeared in our collection.
- We sorted these lists in descending order, extracted top 10 from each of the lists and represented them in tables along with their counts as shown in Table 3.

| Top 10 hashtags | hashtags count |
|---|---|
| Ryanair | 487 |
| ryanaircancellations | 175 |
| ryanair | 80 |
| bustbyfriday | 61 |
| gameover | 56 |
| marian | 38 |
| news | 28 |
| travel | 27 |
| RyanairCancellations | 24 |
| Ryanaircancelledflights | 22 |

| Top 10 user_mentions | user_mentions count |
|---|---|
| Ryanair | 2491 |
| LeaveEUOfficial | 249 |
| facua | 157 |
| controladores | 133 |
| A4Europe | 98 |
| Femi_Sorry | 76 |
| MinutemanItaly | 69 |
| Independent | 67 |
| thecarolemalone | 52 |
| SimonCalder | 50 |

**Table 3: Table of top 10 hashtags and a table of top 10 user mentions**

## 3.2. GETTING "ALL" FRIENDS AND "ALL" FOLLOWERS OF A POPULAR USER IN TWITTER

For this problem, we chose Katy Perry as the popular twitter user. Katy Perry is a famous American singer, songwriter. She has 104 million followers on twitter, which makes her the most followed person on twitter [2]. The day when she became the first person reaching 100 million followers, Twitter posted a video compilation on their website of her historical tweets with a message that read “Today, we #WITNESS history.” [3].

### 3.2.1. Data Collection

Twitter provides two different types of API to developers, the REST API and Streaming API which we used to obtain data in the first question. For this specific situation, we chose to use the method under the category of REST API, with which one can easily get the ID list of both friends and followers of a particular user provided his/her screen name. The method is even more powerful when used in conjunction with another method called lookup, which allows the conversion of user IDs into full user objects in bulk [4]. This is what we used for returning the corresponding user screen name with user IDs. However, due to the large number of followers

she has, it was not possible get to all the ID and screen names of all of her followers as we encountered “Rate Limit Exceeded” error.

The process when extracting user ID resembles the process when you are viewing the information on webpages and turning the pages manually. At most 5000 IDs could be printed on the same page and you need to move the page to the bottom and click some button to give Twitter an information that you already finished reading this page and ready to move to next page. The most common error we encountered was that Twitter tends to block the connection for a while when it believes the crawl rate is too fast, even though we already set to stop 15 seconds every time after we finished a page, and we needed to retry after 15 minutes breaks. To get as many IDs as possible, we built a more robust function to access the Twitter data by pre-treating some common error messages referring to the sample code. Also, each time after we finished a page, we saved the IDs we acquired into a JSON file before continuing to next page. This design guaranteed that we would not lose all our data when the service disconnected accidentally and could save memory when the dataset was too large.

### 3.2.2. Getting "ALL" friends and "ALL" followers of a popular user in twitter

We collected the IDs of her 205 friends and 200,000 followers out of the 104 million followers she currently has. Although we collected 200,000 followers of Katy Perry, it is still a small portion compared to her followers’ number in total. We randomly chose 20 out of each of her friends and followers. Table 4 shows a table we plotted for 20 of her friends with their ID numbers and screen names and another table we plotted for 20 of her followers with their ID numbers and screen names.

20 of Katy Perry's Friends:

| id | screen_name |
|---|---|
| 11348282 | NASA |
| 346276932 | TaraBrach |
| 19725644 | neiltyson |
| 52544275 | IvankaTrump |
| 1652541 | Reuters |
| 51241574 | AP |
| 29450962 | repjohnlewis |
| 63302020 | JordanPeele |
| 181572333 | chancetherapper |
| 757303975 | ChelseaClinton |
| 21288052 | zanelowe |
| 88975905 | shanesmith30 |
| 3327720838 | brielarson |
| 82455213 | RuPaul |
| 39364684 | chrissyteigen |
| 1923827275 | SkipMarley |
| 29417304 | deray |
| 531561605 | AmericaFerrera |
| 796974685288157184 | iwillharness |
| 784152188092190720 | kpcollections |

20 of Katy Perry's Followers:

| id | screen_name |
|---|---|
| 908693273559961600 | sartre6990 |
| 908693184456347648 | ZubairC05824659 |
| 908692829353926657 | MissKagame |
| 908693389213761536 | Darmawan126 |
| 908693198469582853 | gigicarpenter51 |
| 908693353138548736 | EmilyLe21361259 |
| 4619960298 | purnomoarif313 |
| 908692967359053825 | Virgini12060771 |
| 908693190089175040 | Crisel94082929 |
| 886281281759215616 | AlperCiftci8 |
| 900599865096273920 | MagarVaishnavi |
| 906158287783174148 | lisa128911 |
| 1917952952 | Tymeshionna |
| 908692791210868736 | DullaVicky |
| 908693240655867904 | MiriamM61123558 |
| 908693154932523009 | Sanjayj76299545 |
| 908692677058813953 | alejavelez2785 |
| 908540579398639616 | REBiRTH1027 |
| 908690321009201153 | LouLouOttawa |
| 908692577335070720 | Y4gCR8ozDH9GNXs |

**Table 4: Table of 20 of Perry’s friends and a Table of 20 of Perry’s follower**

### 3.2.3. Compute the mutual friends within the two groups, i.e., the users who are in both friend list and follower list, plot their ID numbers and screen names in a table

In this section, we computed the mutual friends within the two groups, *i.e.* the users who are both in the friends and the follower's lists. However, the result we got was empty for the intersection of the friends and followers list. This is most probably due to the number of her friends, which is 205, being so small compared to the number of her followers which is around 100 million. Even if there were mutual ID's in both lists, since the followers data sample we collected was only 0.19% of her total followers, we only had a slight chance to have those who are exactly in both lists and it seems the data we collected did not have any mutual ID's, if any. Hence, we got an empty collection for this part, as shown in Table 5.

```
Mutual Friends:

+----+-------------+
| id | screen_name |
+----+-------------+
+----+-------------+
```

**Table 5: Table of Katy Perry's mutual friends**

## 3.3. BUSINESS QUESTION:

In this section, we describe the business question and discuss in detail about the analysis performed for each of the business question.

**Question 1: Was this move by Ryanair, that is discussed at the beginning of this report, was a good business strategy?**

To answer this question, we did a sentiment analysis on all of the tweets that we collected related to Ryanair. For the sentimental analysis, we used NLTK library, and we did the following:

- We generated word libraries of both positive words and negative words.
- We then analyzed each of the tweets by comparing all the words with the words in the word libraries of positive and negative words.
- Finally, we calculated the percentage of each type; positive, negative, and neutral words, by counting the total number of words of each kind and plotted the results for the sentimental analysis as a pie graph as shown in Figure 1.

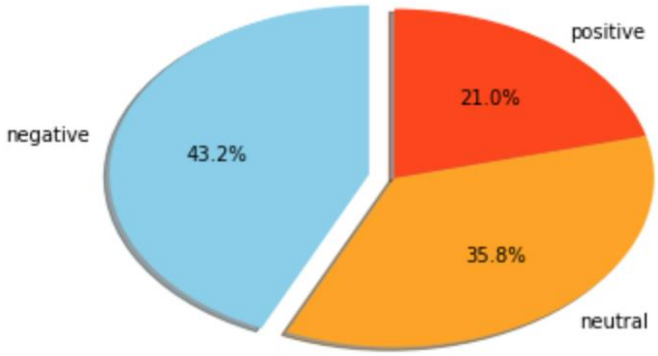


**Figure 1: Pie chart result of the sentiment analysis**

**Question 2: Which locations were most affected by this situation?**

To answer this question, we looked up for the various locations from where these tweets came and plotted histogram of the countries in our dataset. When analyzing countries from which people sent their tweets about Ryanair, we need to take several factors, such as the popularity of airports, location, or even the popularity of twitter in the local area, into account. When flights suddenly get cancelled at airports that have more traffic, it is possible that the number of passengers who were planning to travel from those airports that got affected by the cancellations would be more. We analyzed the twitter data to see if this was reflected by the number of tweets from these locations. UK, Netherlands, France, Germany have the busiest airports in Europe [5] and we also see some of these countries come up in our analysis as shown in Figure 2.

Furthermore, the Dublin city is ranked the highest among other cities (see Table 6) in our analysis, given that Ryanair is an Irish airline this could be expected, however the Dublin Airport is only ranked the 15th busiest airport in Europe [6]. Of course, we cannot make any conclusions from our location analysis since most of the tweets in our dataset is missing the location information. However, it would be interesting to do more analysis to see if more passengers were affected in Dublin than other cities, provided a larger dataset.

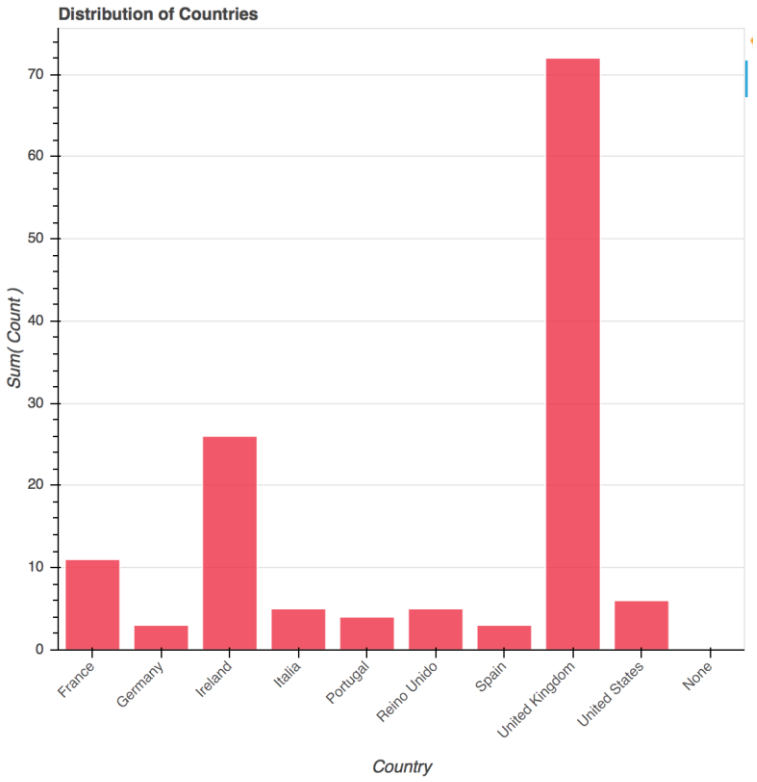


**Figure 2: Distribution of Countries**

Top 20 Cities:

| City | Count |
|---|---|
| None | 5840 |
| Dublin City | 11 |
| East | 5 |
| London Stansted Airport - STN | 4 |
| Sheffield | 4 |
| Manchester | 3 |
| Louth | 3 |
| Lux | 3 |
| Northern Ireland | 3 |
| Camden Town | 2 |
| South Dublin | 2 |
| Stone | 2 |
| Dorking | 2 |
| Lewes | 2 |
| Hetton | 2 |
| Hilltown | 2 |
| Brent | 2 |
| Meath | 2 |
| Edinburgh | 2 |
| Blagnac | 2 |

**Table 6: Table of the top 20 cities**

**Question 3: How could Twitter data help a company decide how to spend its resources?**

We only had 6,000 tweets that were collected in one day. One should, in reality, collect more data preferably until the end of October, until which the cancellations will continue, to analyze the data to get more concrete results. From the results we have, we believe this was not a good business strategy and this will probably cost Ryanair to lose some of its future business, in addition to the liability in the compensation payouts. Hiring more employees might have cost them less. Ryanair is obliged to pay €250 cash compensation for each passenger on a cancelled flight of up to 1,500 km, rising to €400 for longer flights [1]. If 400,000 passengers have their flights cancelled, the total liability in compensation payouts is likely to be around £100 million, though not all who got impacted by this might claim [1]. Their reputation is going to be needed to be restored as the trust element has most likely been lost among the customers. They might need to allocate some of their advertising money to have alluring deals for the locations where they lost the most business to get customers back.

## 4. CONCLUSION

We have collected data on tweets containing the word "Ryanair" using the Streaming API and collected data on the account of Katy Perry for her followers and friends using the REST API. For the Ryanair data, we first cleaned the data and made it ready for our analysis. We were able to find the top 30 words, top 10 tweets and top10 hashtags and user mentions, and ranked the countries and the cities the tweets were sent from to help identify which locations were affected the most. We also did a sentiment analysis on the data, in order to identify and categorize user opinions. Using these results, we answered couple questions regarding business and came up with solutions as suggestions to what Ryanair could do to reverse the loss that will result from this action they took. We also analyzed the data we collected from Katy Perry's account regarding her followers and friends and after cleaning the data, we were able to extract 20 of her followers and 20 of her friends and found out that in our collection of data, we were not able to capture any mutual friends in both her followers and friends lists.

## REFERENCES


[1] “Ryanair: Among the 400,000 passengers whose flights are grounded? Know your rights.” The Independent. September 16, 2017. https://www.yahoo.com/news/ryanair-among-400-000-passengers-112102713.html

[2] Top 100 Most Followed Users on Twitter. https://twittercounter.com/pages/100

[3] O’Connor, Roisin. "Katy Perry is first to reach 100m Twitter followers." The Independent. June 17, 2017. http://www.independent.co.uk/arts-entertainment/music/news/katy-perry-twitter-followers-100-million-first-obama-taylor-swift-rihanna-justin-bieber-a7794726.html

[4] GET users/lookup. Twitter Developer Documentation. https://dev.twitter.com/rest/reference/get/users/lookup

[5] Gulliver, “Busiest airports in Europe. The Economist.” February 20, 2017. https://www.economist.com/blogs/gulliver/2017/02/lure-london

[6] “List of the busiest airports in Europe." Wikipedia. https://en.wikipedia.org/wiki/List_of_the_busiest_airports_in_Europe