Trends detection on Twitter

Anh Nghia Khau **Sandra Djambazovska** **Camille Renner**
{anh.khau,sandra.djambazovska,camille.renner}@epfl.ch

Abstract

Through the years, Twitter has grown in one of principal sources of trends in the world. The goal of this project is to detect trends in tweets. In order to achieve this, we extract the topics using LDA jointly with the finding and discussion of the top trends per month in 2014.

1 Introduction

Twitter is a social networking service which users can read and post short messages called tweets. In 2017, this platform has reached 330 millions of monthly active users (Statista, 2017). Tweets have a restricted number of characters and after being posted they can then be retweeted by other users. Simple and intuitive, this major microblogging platform became a mirror of current trends as well as a generator of new trends. Social media is a big part of our lives. It brings people together, they share their lives, personal information and views.

For these reasons, dynamics of trends on Twitter will be explored in this paper. The following research question will be investigated: "Which topics are the most tweeted (retweeted) and spread the fastest in time?".

The paper is structured as follows: Section 2 presents related works, Sections 3 briefly describe the dataset collection and composition while Section 4 details the methods used in this project. Then, Section 5 shows the results obtained and Section 6 provides a discussion and conclusion of this study.

2 State-of-the-art

The detection of trends on social networks and especially Twitter is analyzed in numerous studies. For many of these studies, the real-time aspect is a key challenge for trend detection (Benhardus and Kalita, 2013; Mathioudakis and Koudas, 2010). On another hand, several different methods are applied in order to distinguish topics leading the trends. Aiello et al.(2013) compared three traditional algorithms with four novels approaches for topic extraction and trend detection. While they proved that a new technique using n-grams and tf-idf, metric that introduces time to the classic tf-idf score, topic ranking performed best, they also showed that Latent Dirichlet Allocation (LDA) performs well on focused events. Moreover, their analysis found that topic aggregation of tweets could improve the LDA algorithm results while time aggregation did not have an effect. Ultimately, trend detection could be associated with sentiment analyses in order to assess the user's position towards specific trends (Salas et al., 2016).

Given the time and depth of this course, the LDA method was chosen for our analysis.

3 Dataset description

In this project, data collection through Twitter API was not needed as a Twitter dataset was given. Originally, this dataset contains roughly 20 billions tweets in many different languages: English, French, German, Spanish and Dutch. Each data entry has five fields: language, id, username of the user as well as the date of publication of the tweet and its content. By cleaning the dataset, only entries with five complete fields were kept. Moreover, in order to decrease the size of the dataset, only tweets from the year 2014 and for which the language of the user was English, were selected (See Section 4.1). The final dataset contains 1.8 billions tweets (See Section 5).

4 Methods

In this section, the methods used in the topic detection pipeline will be defined. Starting with describing data cleaning and pre-processing, creating the TF-IDF matrix and LDA algorithms. Eventually, the detection in time will be detailed.

Despite the fact that hashtags reveal a lot of useful information about topics and trends and are very present in a Twitter dataset, we decided not to take them into account because of the difficulties to process them as explained in Section 4.2.

4.1 Data cleaning and pre-processing

The dataset we have is dirty and has a lot of noise. So, in order to be able to perform meaningful analysis it is crucial to spend time on data cleaning and pre-processing. To do so, first of all, entries missing one or more components were removed. Then, only tweets with English as user's language were selected. To reduce the noise in the remaining tweets, tokenization and filtering techniques were used:

- **Tokenization**: As the size of tweets is limited, raw posts can include abbreviations or typos. This method enables to map tweets into a sequence of tokens, i.e. words, while removing punctuation, emoji and url. This is a pre-requisite for choosing indexing terms. Here, we were faced with the problem of compound words (e.g. credit card). Nevertheless, we decided to leave them as they are.

- **Filtering**: Tweets contain a lot of irrelevant or meaningless words, like low frequency words or stopwords. Therefore, we are going to exclude them from indexing.

    1. **Stopwords**: Stopwords are widely used on Twitter and therefore have a high frequency but they are meaningless. For example, words like a, the, that, this, etc are not informative. Furthermore, since the data is extremely noisy, we had to combine two different lists of stopwords: the list of stopwords for the English language and the one for stopwords on Twitter. Also, removing this stopwords will help us reducing significantly the number of words to be considered for the semantic analysis, which will result in less computation time. But since the tweet is usually short, it is possible that a sentence like 'to go or not to go' will be entirely removed.
    2. **Lemmatization**: This technique reduces word variations to the root of the word (e.g. walking, walks, walked $\Longrightarrow$ walk). Lemmatization reduces index size and increases information value of each indexing term. For this purpose, we use a third party library `nltk` in which we first find PoS(Part-of-speech) tagging of each word and then maps it to the root word. Finding PoS tags is crucial because the lemmatization method has default `pos` argument set to n(noun): `lemmatize`(are) = are while `lemmatize`(are,pos='v') = be.
    3. **Frequencies**: By removing stopwords, we exclude words that have high frequencies. In the opposite case, we also decided to remove all words that have low frequencies by passing minDF (the fraction of document that a term must appear in, in order to be included in the vocabulary) as an argument when creating the Term Frequency (TF) matrix using pyspark.ml.feature.CountVectorizer object.

4.2 Hashtags frequency

As mentioned above, hashtags carry a lot of important information and it can be very useful for the semantic analysis. However we decided not to take them into account for topic modeling because of the difficulty to split them into tokens (words). For example, the process splits the hashtag `#ilikeitwhen` into "i like it when". Nevertheless, it is interesting to discover which are the hottest hashtags in 2014 (See Section 5).

4.3 TF-IDF

Constructing the term frequency-inverse document frequency, TF-IDF, matrix enables to sort out the most discriminant words in terms of topic. Indeed, it converts the bag-of-words matrix into a weighted one that emphasizes on important words. The IDF represents the inverse document frequency (information content of an event??) and the score for one word is calculated with the following formula:

$$idf(w) = -\log(docfreq(w)/N) \qquad (1)$$

The variable $docfreq(w)$ represents the number of tweets that contain the word w in the whole dataset whereas N is the total number of tweets, in our case $N = 1.8billion$. The variable $tf(w,t)$ depicts the frequency of the word w in the tweet

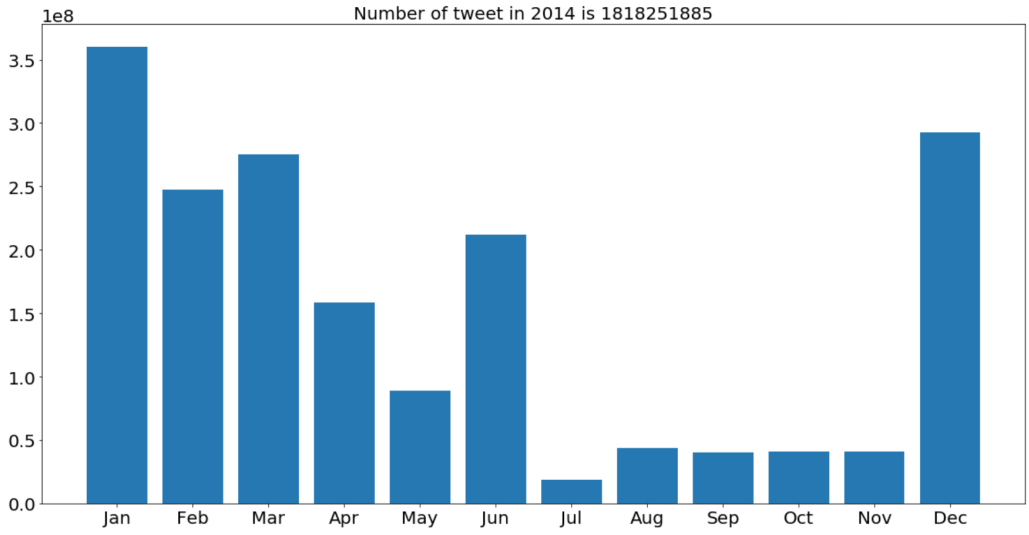

Figure 1: Distribution of tweets in 2014 in our dataset

t and is used to construct the elements of the TF-IDF matrix (See equation 2).

$$TF\text{-}IDF(w,t) = tf(w,t) * idf(w) \quad (2)$$

Thus, high weight in tf-idf is reached by high term frequency in the given tweet and a low document frequency of the term in the whole collection of tweets.

4.4 Topic modeling using LDA

After having done the TF-IDF of our dataset, the Latent Dirichlet Allocation (LDA) is a probabilistic method applied for topic modeling. This unsupervised method takes as inputs the TF-IDF matrix and the number of topics desired, then it gives as output the topics and the distribution of topics over each tweet. Each topic is contributed by a set of terms/tweets that are related to each other.

Using LDA to infer the semantic distribution of our dataset, we are confronted with many challenges that we are going to discuss now:

- **Size of the tf-idf matrix**: First of all we are faced with the problem of big data. Precisely, if we consider all the tweets (1.8 billion tweets and 113 million terms), it will take a long time for the computation. To make it simple, only terms that appear in at least 1000 tweets are considered (only 127k terms are selected) and we take only a subset of those tweets (100k) for the computation.

- **Parameters**: In this project, since we don't have a lot of time and resources, we fixed k(number of topics) = 20, optimizer = 'online', without tuning documents concentration (α) and topics concentration (β).

- **Short tweets**: Another problem is that normally LDA works well on structured and long documents while tweets are extremely noisy, short and unstructured which leads LDA to fail to infer the topics.

4.5 Trend detection in time

Because of the limited time frame for our project, we decided to simplify this task. Our first idea was to detect the time when the trend appears for the first time and follow it until it disappears or when it only appears less then a certain threshold. Instead of doing this, we decided to find the top ten hashtags for each month in 2014. We also thought that it would be interesting to see who are the most popular users on Twitter. We defined popularity as the most retweeted user.

5 Results and findings

5.1 Preliminary dataset analysis

The first step consists of having a general overview of our dataset. As it is observable in Figure 1, the distribution of tweets over the year 2014 is not constant in our dataset. This could be because of our sampling or in a least measure because of the natural fluctuation of tweets depending on trends. Furthermore, the most frequent hashtags are represented in the Table 1 below. Their frequency

distribution can be seen on the Figure 2 in the Appendix.

Top 10 Hashtags		
gameinsight androidgames android, ipadgames, Android	KCA MTVStars porn	openfollow RT

Table 1: Most frequent hashtags in the whole year

In this table, hashtags are divided according to three different categories. The first one contains the most frequent term `#gameinsight` which was automatically generated when users reported their gaming performance alongside the others in this category such as `#androidgames` or `#ipadgames`. In the third one, hashtags inherent of Twitter system are present, for example `RT` stands for retweeted. Only the second column could reflect trends such as `#KCA` for Kids Choice Award, an american ceremony taking place in March or April, `#MTVStars` which elects on social media the best artist of the year and finally, a surprising `#porn`.

5.2 Trend detection in time

Over time, we can see that some of the trends stay for a long period in the top ten like `#gameinsight` or `#android` . But some appear just for a short period like some events or shows (e.g. `#KCA`). A couple of months after the event or show is over people stop talking about it. In conclusion, there are some constant trends that stay for a long period but most of them depend on the current events and happenings. We also have trends that appear each year or each four years, for example elections, sports event (`#WorldCup`), award ceremonies etc. This is visible in the Figures 3 and 4 available in the appendix, where the `#KCA` explodes in March(during the Kids Choice Awards), while the `#gameinsight` stayed constant.

Concerning the most popular users, without surprise they are famous entertainers, most of them from the music business and popular among teenagers. An example for that is Justin Bieber or the members of One Direction.

5.3 Topic modeling results

Topic modeling method using LDA produces mixed results. Indeed, out of the twenty topics, some are good whereas others are just a mix of non-related words. The Table 2 shows five examples of extracted topics.

Example of some topics	
topic 1	Happy, New, amazing, Heres, fireworks, Lets, 2014, 12, cool, midnight
topic 2	followers, 3, stats, unfollowers, 15, 14, far, ugly, Really, fireworks
topic 3	tell, said, xx, coming, hell, 4, maybe, ld, wear
topic 4	getting, Oh, team, nah, selfies, chance, chill, Lee, youve, dance
topic 5	yes, face, actually, leave, babe, ball, drop, shot, dick, fuckin

Table 2: Topics detection with LDA

While topic 1 is clearly representing the New Year Eve corpus, topic 2 and 4 could be representative, respectively, of Twitter statistic topic and the corpus around fun activities. On the other hand, topics 3 and 4 are not really sorting a clear topic. Moreover, despite the cleaning and preprocessing steps, there are still a lot of words without real meaning.

These results could benefit from deeper cleaning and analysing but because of the time constrain and cluster overload, we stopped at this point.

Overall, we discovered that some trends such as Kids Choice Award or New Year Eve are short in time but intense whereas others like gameinsight or Twitter statistics are present all year long.

6 Conclusion

Through this project, we faced the real struggle of handling a huge amount of data extracted from social media. This kind of data are often noisy and dirty and therefore require substantial cleaning and pre-processing steps. Without taking hashtags in account, topics, which are more or less focused, were extracted from our dataset. On another hand, the analysis of hashtags provided interesting results for trends. From our results, we might even deduct that users in our dataset were probably young people. This study could be extend to the spacial dimension in order to see where trends are created and where do they spread.

References

Aiello, L.M., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., Gker, A., Kompatsiaris, I., Jaimes, A., 2013. Sensing Trending Topics in Twitter *IEEE Transactions on Multimedia.* Vol 15, 1268 - 1282.

Benhardus, J. and Kalita, J., 2013. Streaming trend detection in Twitter. *Int. J. Web Based Communities* Vol. 9, No. 1, pp.122–139.

Hoffman, M. D., Bach, F., and Blei, D. M., 2010. Online Learning for Latent Dirichlet Allocation. *Advances in Neural Information processing systems.* pp.856–864.

Mathioudakis, M. and Koudas, N., 2010. TwitterMonitor: trend detection over the twitter stream. *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data.* pp. 1155-1158.

Milioris, D., 2015. Trend Detection and Information Propagation in Dynamic Social Networks. Document and Text Processing. Ecole Polytechnique.

Milioris, D. and Jacquet, P., 2015. Topic detection and compressed classification in Twitter. *23rd European Signal Processing Conference (EUSIPCO)* Nice, 2015, pp. 1905-1909.

Salas-Zrate M..P., Medina-Moreira J., lvarez-Sagubay P.J., Lagos-Ortiz K., Paredes-Valverde M.A., Valencia-Garca R, 2016. Sentiment Analysis and Trend Detection in Twitter. (eds) Technologies and Innovation. CITI 2016. *Communications in Computer and Information Science.* Vol 658. Springer, Cham.

Statista: the statistics portal. Twitter: number of monthly active users 2010-2017. *(Visited on December 15, 2017).* Retrieved from: `https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/`.

A Appendix

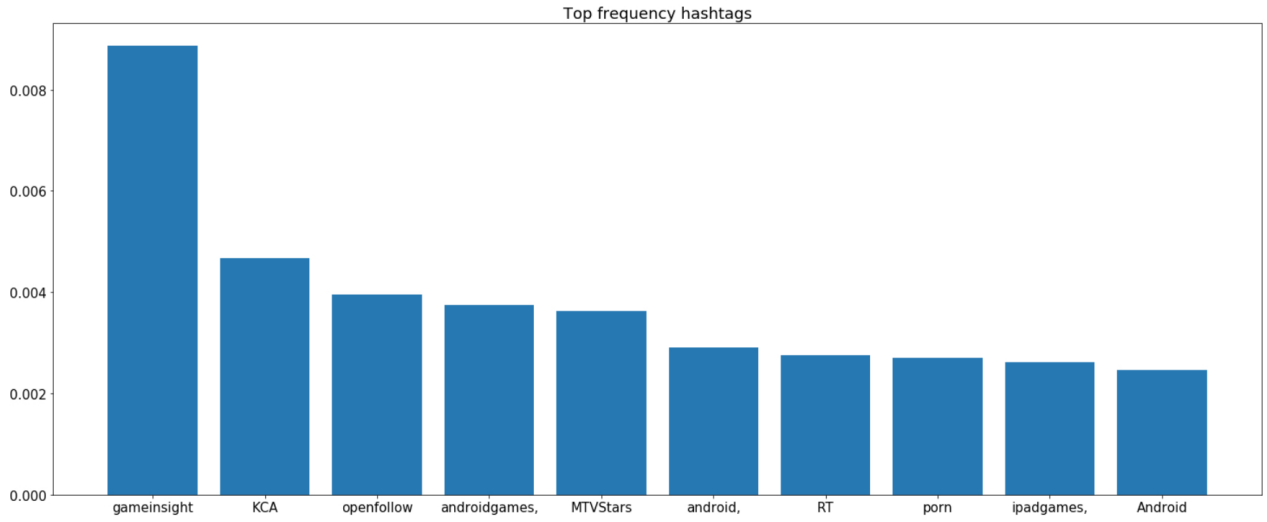

Figure 2: Most frequent hashtags in 2014 in our dataset

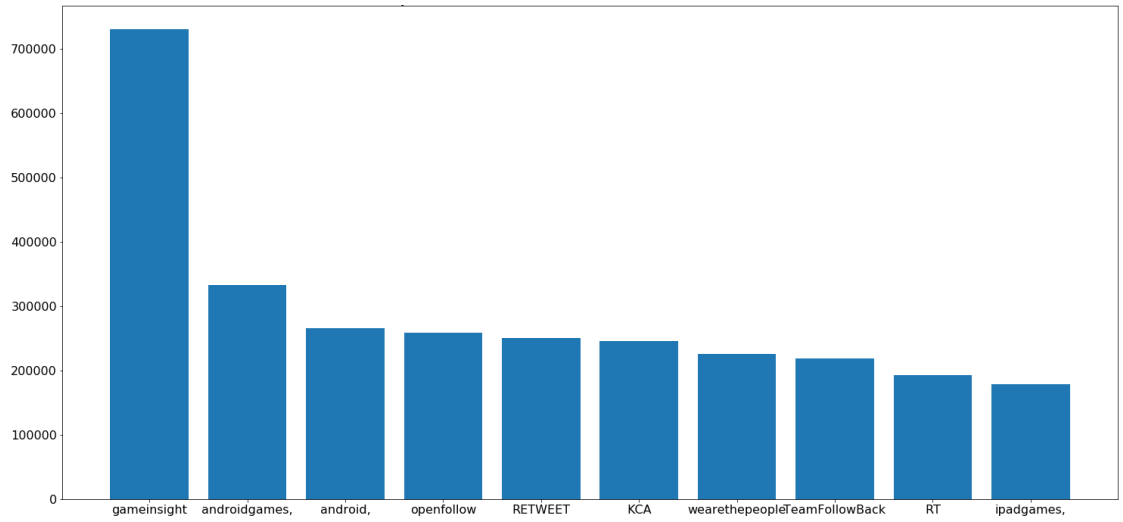

Figure 3: Top 10 hashtags in February

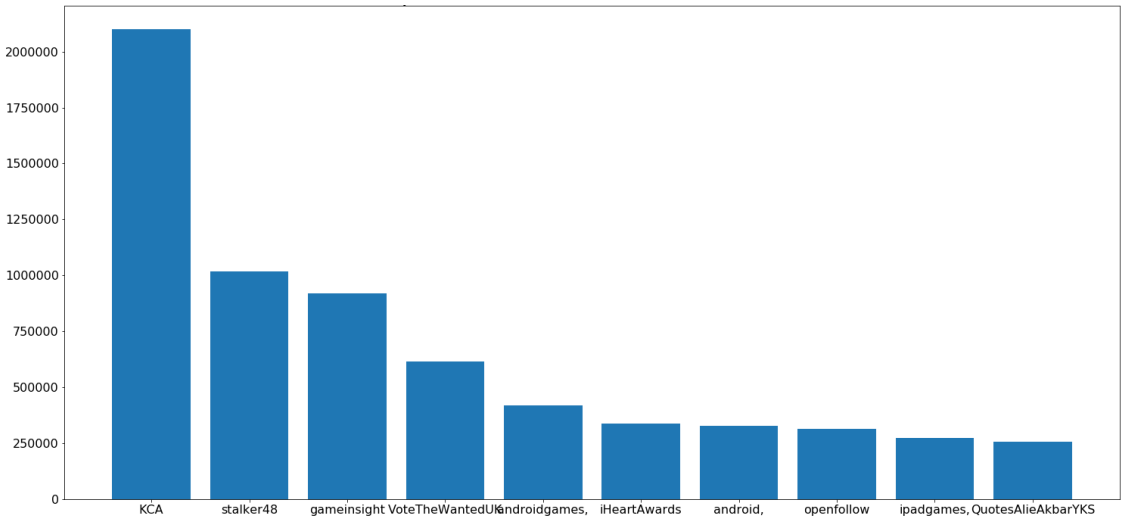

Figure 4: Top 10 hashtags in March