MEGHAN AND HARRY'S INTERVIEW: ANALYSIS OF USERS' BEHAVIOUR ON TWITTER

Alessandro Fossati, *University Bicocca of Milan, Faculty of Data Science*

May 16, 2021

Abstract

Events concerning The Royal Family are always discussed, not only by Uk's Tabloids, but also by people all over the world. In Social Media Era, even more people have its say about these events, and so, it has been very interesting to know what people think about Meghan and Harry's case. They decide to leave Buckingham Palace to go and live in US, breaking ties with Royal Family and breaking their work contracts indeed, as a result of some big problems encountered in everyday life, which could have caused serious damages to Meghan's person, and to Harry and Archie (their son) consequently. In fact, the duchess of Sussex has declared to have been in a deep state of anxiety and depression for long time, because of the pressure that she has to take. In this way, there was a clash between the dukes of Sussex and the Royal Family. Oprah Winfrey, US television presenter, in agreement with Meghan and Harry, interviewed them to shed light on their case. A lot of topics have been tackled in Oprah's interview, like depression, suicide, racism and much more. This interview was broadcast on US's TVs in the early 8th March 2021 and was broadcast on UK's TV in the late 8th March 2021. In this paper we want to show what people think about some topics presented during Oprah's Interview by analyzing users' behaviour on Twitter platform mainly in term of Sentiment Analysis, Emotion Analysis and Emoji analysis.

1 Introduction

To analyze which is people's behaviour about an event and to analyze the number of people interested in it, could be very useful in order to understand the size and the relevance of that event. In this way we aim to discover people's emotions and feelings about some topics of Meghan and Harry's case. In such a big case we can find a lot of topics to discuss, but we decide to analyze specific topics after having explored data collected from Twitter through an interactive data visualization on Tableau, which will be presented in detail in this report. In addition to explorative aim, some research questions that led us to produce Tableau visualization was: *which is the trend of tweets categorized as positive, negative or neutral during our data collection? Which is the most commented interview? Which are the most used words during the interviews?* Answering these questions we found other interesting topics and research questions to answer. *Which were users' emotions and sentiment related to Meghan Markle during both interviews? Which were users' emotions and sentiment in relation to racism topic, which comes from Archie's case? In which way retweets could influence the values of sentiment score obtained?* These three research questions are the core of the project. In relation to this third question, we decide to *explore retweet networks of UK and US's interview*, analyzing users' behaviour in the usage of retweet functionality. One last funny but not obvious research question is related to *looking for the top emojis used in both interviews.*

The answers to these questions will be provided in this paper, which is first of all composed by a section where will be given details about Twitter data collection. Afterwards there will be a section dedicated to technical instruments in order to obtain results of sentiment analysis, emotion analysis, emoji analysis and retweets analysis. In the following section will be shown Tableau interactive dashboard, where research questions arise. After that will be found a sequence of sections concerning the results obtained in relation to the research questions presented. Finally, will be presented conclusions about this research project.

2 Data Collection

Data collection is carried out using *Tweepy*, a Python Library which allows to access *Twitter Developer APIs*[1]. The collection is performed in batch, and refers to 8th March 2021, date of US and UK's broadcasting of Oprah's interview. Data collection is a search by hashtags which relate to the interview, like *#OprahMeganHarry*, *#MeganandHarryonOprah*, *#meghanandharry*. Hashtags are selected manually exploring Twitter, but also checking for hashtags in trend for the specific day[2]. In this way we obtain a sample of approximately 95k tweets with their corresponding data and metadata used for the analysis, contained in some fields like:

- *created_at*: field which refers to the date of publication of the tweet.
- *id*: field which refers to the unique identifier of the tweet.
- *text*: field which refers to the text of the tweet.
- *retweeted_status*: field which presence indicates that a tweet is a retweet. This field is definitely significant in our analysis, given that retweets are present in all the analysis carried out.
- *lang*: field which refers to the language in which the tweet is written. We decide to keep only tweets written in English, which are approximately 89k, almost all of the tweets collected.
- *user.screen_name*: field which refers to the scree name of the user who publishes the tweet.

All tweets are collected in a DataFrame *Pandas* and are exported to CSV data format, in order to perform analysis successively.

[1] 1.
[2] 2.

3 Technical instruments For analysis

In this section of the paper we want to discuss in details technical instruments we used in order to carry on analysis. In this way will be shown five subsections respectively related to *Text Preprocessing*, *Sentiment Analysis*, *Emotion Analysis*, *Emoji Analysis*, and *Retweets Analysis*.

3.1 Text Preprocessing

In order to produce more truthful analysis is fair to provide text preprocessing of tweets. In this way *nltk* python library allows us to obtain tokenized and clean tweet's text removing stopwords like articles, pronouns and conjunctions. Moreover, is fair to remove unnecessary punctuations or various symbols. Next step is lemmatization of text, powered by *WordNet*, a very important operation if you decide to provide a lexicon-based sentiment analysis. Lemmatization carries words to their lemma, in order to have a better match with dictionaries or lexicons in sentiment analysis' phase.

3.2 Sentiment Analysis

Sentiment analysis is a sub-field of *NLP* that tries to identify and extract opinions within a given text. In this case, sentiment analysis is powered by *VADER*[3], a rule-based model created by C.J.Hutto. VADER is used for lexicon-based sentiment analysis, particularly suitable for social media analysis. Principal core of lexicon-based sentiment analysis is the fact of leading back tokens of text to their original lemmas, as anticipated. In this way is possible to search and match in the lexicon or in the dictionary the related sentiment score. In this analysis is used *vader-lexicon*. In order to compute a sentiment score, after a text preprocessing phase, for each tweet in our sample VADER provides a *compound score* that is a weighted mean of the normalized polarity score (between -1 and 1) assigned to each word in the text of a tweet. In this way every tweet will obtain a compound score, a normalized weighted sentiment score (between -1 and 1).
In this analysis we are also interested in the classification of each tweet as positive, negative or neutral, relying on compound score. A text of a tweet is classified as negative if its compound score is under -0.05, as positive

[3] 3.

if its compound score is over +0.05, and neutral otherwise. However, for this project is interesting to obtain sentiment scores related to some specific periods as we will see later. For example, could be interesting to analyze sentiment score during UK interview, or during US interview. So is useful to compute an average value of the compound scores of the tweets which belong to the specific subset considered. In order to obtain a more reliable result, could be useful to provide a confidence interval for the mean of the compound scores of a subset. So, we provide a 90% level interval for the *Gaussian* probability distribution. Finally, in order to visualize sentiment results we use *matplotlib* python library and *Tableau Software*.

3.3 Emotion analysis

Emotion analysis is one of the principal core of the project and is aimed at discovering which emotions users feel in relation to some topic. Emotion analysis is carried on using *NRCLex*[4] python library. NRCLex measures emotional affect from a body of text. Affect dictionary contains approximately *27k words*, and is based on the National Research Council Canada (NRC) affect lexicon and the NLTK library's WordNet synonym sets. In this way every word in tweet's text, if belonging to lexicon, is associated to one or more emotions, which are anticipation, fear, disgust, surprise, sadness, anger, trust and joy. In order to visualize emotions analysis' results has been used pie charts created with MatplotLib python library. Relating to emotion analysis we decide to exclude retweets from the analysis, considering that NRClex associates words in tweet's text with one or more emotions. So, in a retweet, words are not written by users who retweet the post, and so emotion analysis in this case could result a little bit shallow.

3.4 Emoji Analysis

In order to provide *Emoji analysis* we use *emoji* python library, which allows us to detect emojis in tweets' text. Even in this case we decide to exclude retweets from the analysis because if the post retweeted would have not contained an emoji, probably would be equally posted by the user. We finally plot results with matplotlib barcharts, in order to show which emojis were most used by users.

[4] 4.

3.5 Retweets Analysis

The last analysis we provide is *Retweet analysis* because they constitute about 40% of the sample of tweets we got. For both interviews we build retweets' network using *network* python library, and we provide barcharts with matplotlib in order to analyze users' behaviour in using this functionality of Twitter. Retweet functionality is definitely the most used and known action on Twitter, and you can retweet a post with only two taps. In this way you can post something you like or some ideal you share. In *Social Media Analysis* retweets could be very useful, but could also be something that misguide analysis if it is not handled carefully. In this way we decide to carry on our analysis both considering and not considering them, according to the analysis we decide to carry on. For example, as anticipated, retweets have been excluded from Emotion and Emoji analysis, and are central in Sentiment analysis.

4 Explorative Tableau Dashboard

Which is the trend of tweets categorized as positive, negative or neutral during our data collection? Which is the most tweeted interview? Which are the most used words during the interviews? These are the explorative research questions we want to answer with this explorative dashboard produced using Tableau (Fig4.1). In this section will be presented dashboard as static, but you can exploit interactive options following link in *References* section[5].

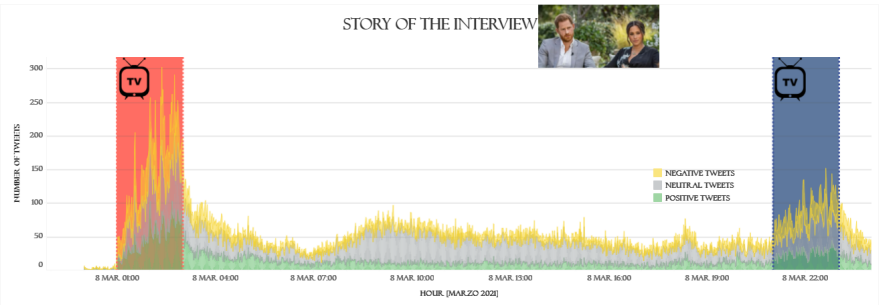

Fig4.1

[5] 5.

In the upper part of this dashboard you can find a time series of the number of tweets collected for every minute of the day. Time and tweets related to US and UK's interviews are reported in red and blue bars respectively. Is clear that the most followed and commented interview is the US one, probably for population size reasons, or maybe because of the topics covered during the interview as we will see.
Moreover, following the legend reported, is possible to have a first look at the size of tweets which are categorized as *negatives, neutrals or positives* with *VADER compound score* of sentiment analysis, with an *area chart* visualization. Is clear how neutrals tweets are dominant when the interviews are not broadcast, while negatives or positive tweets are outnumbered in this period of the day. However, as predicted, they grow up during both interviews because of people is more engaged watching live the interview.

In the lower part of the dashboard (Fig4.2) are shown *cloudwords* related to the interviews (note correspondence of colors among cloudwords and bars). First of all are presented the top 15 words used by users for each interview associated with a more vivid color if they are more used. Moving into the bars, cloudwords will change showing the number of times that the top 15 words used in the interviews appear in the set of tweet (filtered by reweets) related to the minute selected.

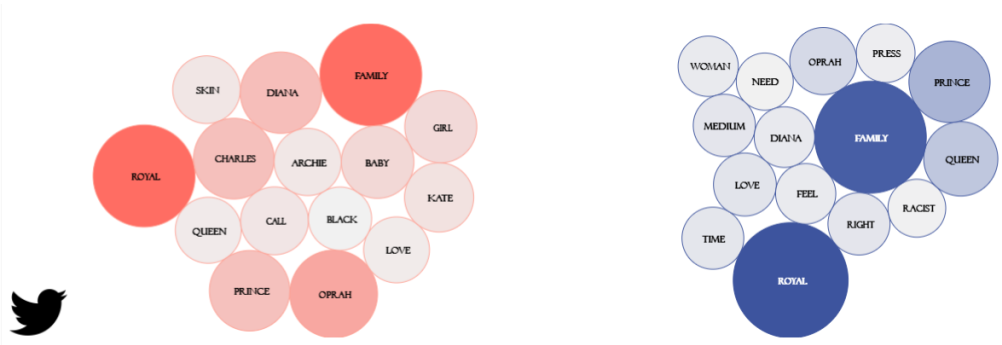

Fig4.2

Positioning on minute 33 after the beginning of the interviews is possible to see that the most used word is definitely *skin*, which refers to Archie's topic (Fig4.3). In fact, Royal Family was scared about the fact that Archie,

Meghan and Harry's son, could have had black skin at his birth. By Royal Family could have been difficult to accept that prince's skin could not have been white.

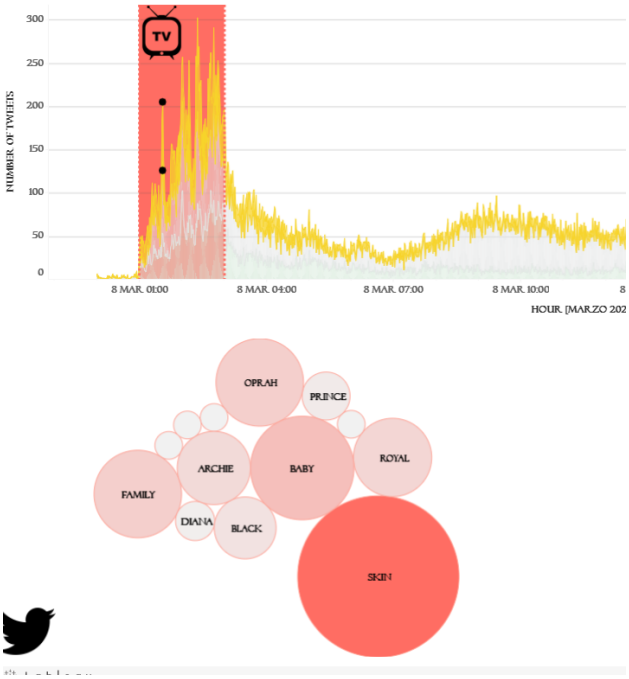

Fig4.3

In this way a new research question comes: which were users' emotions and sentiment in relation to *racism* topic, which comes from Archie's case? Consequently, which were users' emotions and sentiment related to *Meghan Markle* during both interviews?
The words *'Meghan'* or *'Markle'* do not appear in the cloudword because they were too much bulky. In fact, Meghan is definitely the protagonist of the interview, so it goes without saying to include her name and surname in the cloudwords. Nevertheless, along with the topic of racism, what users think about Meghan will be one of the core of the analysis. In this way the first section of the paper will be dedicated to these two topics, passing through Sentiment and Emotion Analysis.

5 Sentiment and Emotion Analysis

In this section, as anticipated will be shown results of sentiment and emotion analysis of some principal topics encountered during Oprah's interview. There will be a focus on *Meghan Markle*, who is definitely the protagonist of the interview. Is interesting to discover which are emotions felt by users in relation with topics faced by Meghan during Oprah's interview. There will also be a focus on *racism topic*. Could be very interesting to discover how users behave on this giant topic. We aim to study sentiment analysis scores and emotion analysis results as measurement and quantification of behaviours and emotions of users. Both analysis of sentiment score include retweets, while both emotions analysis exclude them, for reasons explained in the third section of the paper.

5.1 Focus on Meghan

Analyzing Meghan's tweets, have been applied some filters in order to produce a better analysis. We search in our sample tweets containing words or hashtags which could lead us back with certainty to tweets regarding Meghan, like *meghan, #meghanmarkle, megxit*. We select tweets and retweets as anticipated, coming from both interviews separately, in order to have a comparison in this sense. From our sample of 95k tweets, we obtain 10k tweets regarding Meghan, more than 10% of them. Approximately half of them are retweets, which will be used only for sentiment analysis, which is the first analysis carried on (Fig 5.1.1).

In Fig5.1.1 is possible to understand how retweets could change the results of analysis. Keeping count of retweets you can observe how the difference of sentiment score between US and UK's interviews is statistically different according to Gaussian confidence interval, with UK negative score in according to VADER score. Excluding retweets from analysis we can notice that there is no more statistical evidence about the difference between the result of sentiment score between the two interviews.
Extracting a comment of general order is not easy, even because topics encountered during the interview are strong and hot. In this way, every comment could be biased, because an higher compound score could be associated to a bad word but also to a strong emotion's word. So, at now, we can put in evidence the fact that in this set of tweets, retweeted post during UK's

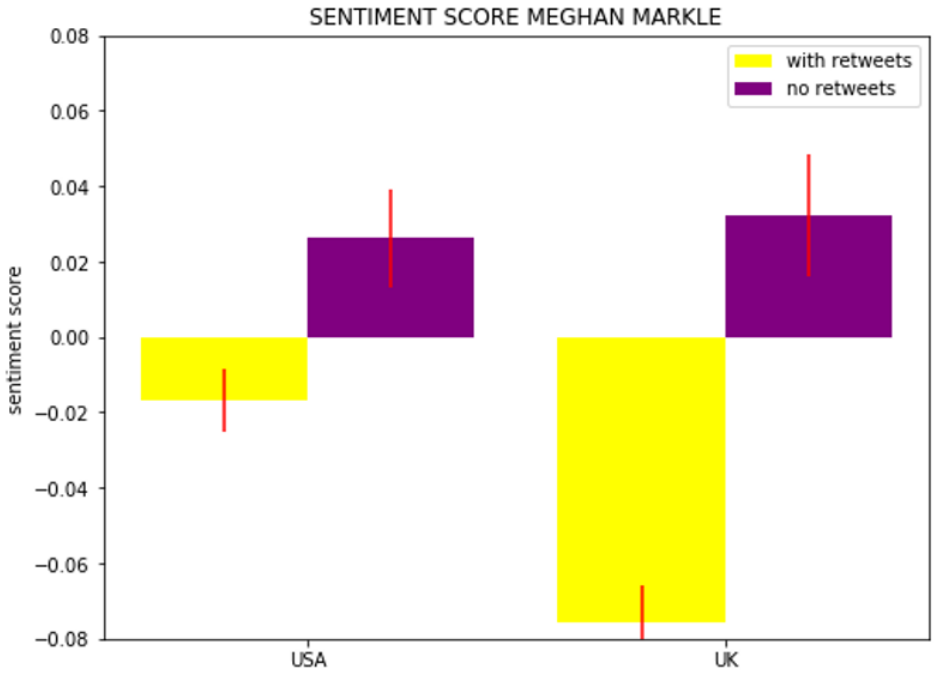

Fig5.1.1

interview could have contained strong words, more than retweeted post of US's interview.
In order to better understand users' behaviour about Meghan is useful to analyze their emotions (Fig 5.1.2).

Pie charts related to USA and to UK are very similar, with prevalence of negative emotions like sadness above all, anger, disgust and fear, which in both cases constitute approximately 60% of emotions collected with NRClex. Prevalence of emotions like sadness or anger are explainable with the topics encountered during Oprah's interview. Oprah talks with Meghan about racism in relation with Archie's facts, talks about depression and suicide in relation with the difficulties encountered by Meghan in Buckingham palace every day's life. In this way people reacts with words which recall negative emotions. Anyway, we can find also positive emotions like joy or anticipation, probably related to the birth of Harry and Meghan's daughter. These results obtained from emotion analysis give an explanation about negative

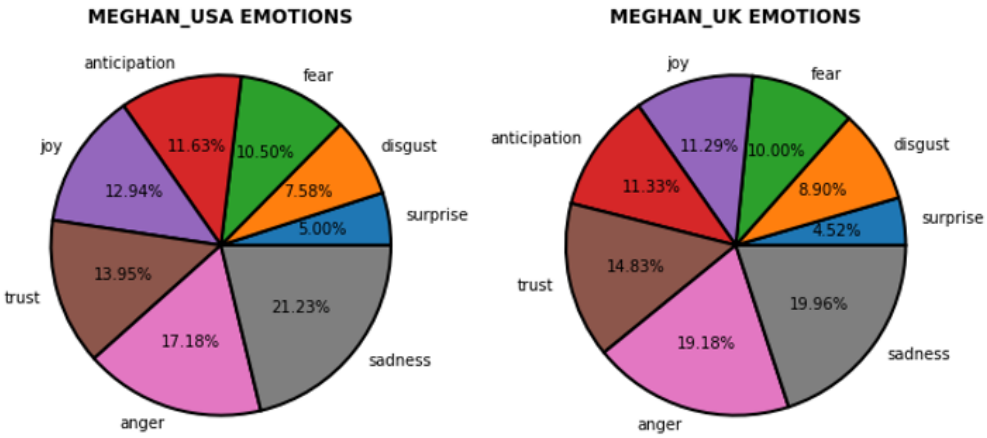

Fig5.1.2

sentiment score obtained during UK's interview.

5.2 Focus on Racism

Even analyzing racism's tweets, have been applied some filters in order to produce a better analysis. We search in our sample tweets containing words or hashtags which could lead us back with certainty to tweets regarding racism, like *#blacklivesmatter, Archie, skin* and so on. We select tweets and retweets coming from all the collection, not only relating to US and UK's interviews as previous. From our sample of 95k tweets, we obtain 22k tweets regarding racism, more than 20% of them. The interesting thing is that only 3500 tweets of this sample are not retweets. This is definitely a topic which users favour to retweet instead of tweeting with their own hand. Let's analyze sentiment score obtained (Fig5.2.1).

In general, as predicted, sentiment scores are massively negative, because of the topic analyzed. Anyway, you can notice statistically significant difference (for Gaussian 90% interval) between score with retweets and score without retweets. This difference could be explained with the different sizes of the two subsamples as anticipated. Another explanation could be that people who is retweeted are in general famous people in the field of Social

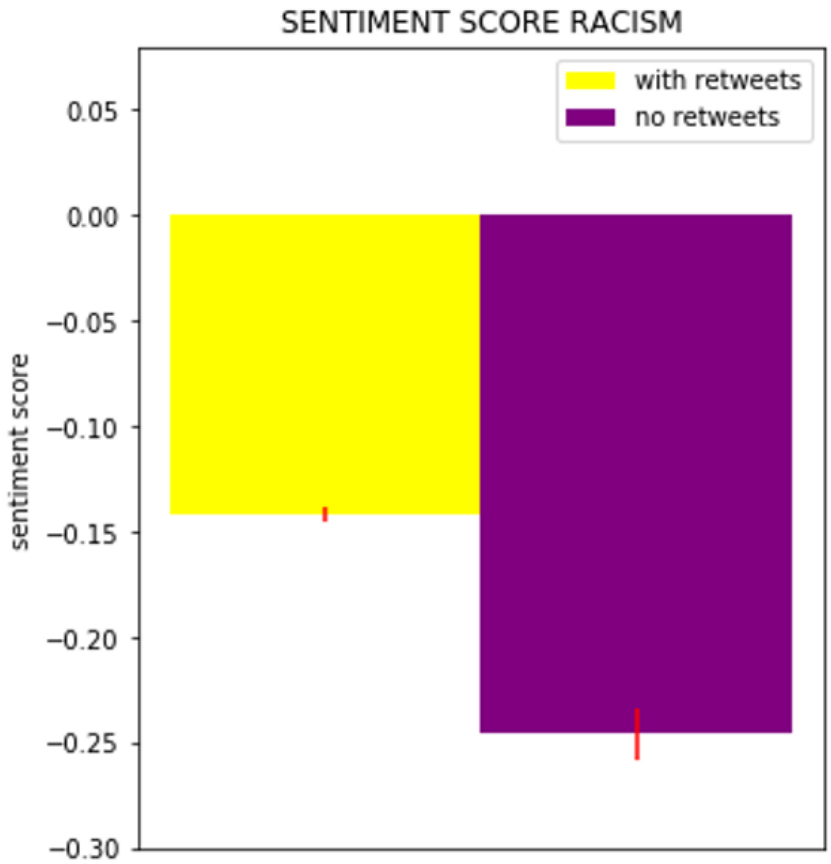

Fig5.2.1

Media, so they have to be moderate in a certain way with the words they use. Anyway, a negative score could come from words belonging to negative emotional sphere, and so previous interpretation of the result could be weak. Is not easy to give an univocal explanation to sentiment scores, because of this topic carries along negativity. In this way we can exploit emotional pie charts and emotion analysis to better understand (Fig5.2.2).

Result of emotion analysis is similar to Meghan's pie charts, demonstrating that the two analysis are linked, and that some tweets are common to both topics. Even in this case negative emotions take about 60% of the pie, with the clear prevalence of sadness emotions. We could have expected an higher value of anger in relation with the topic of racism. Is clear that pos-

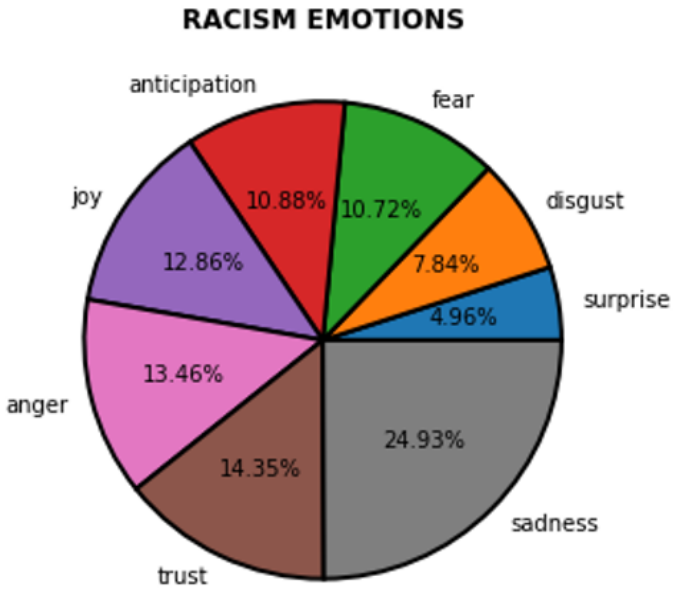

Fig5.2.2

itive emotions are related to the fact that even tweets containing the word *Archie* has been searched for this analysis.

6 Emoji Analysis

In this section of the paper we want to show results concerning *Emoji Analysis*. As anticipated, we search for top 6 Emoji used by users during both US and UK's interviews (Fig6.1), excluding retweets.

We collect approximately 2000 emojis for US's interview and approximately 1300 emojis for UK's interview. Excluding retweets and filtering by hour in order to obtain tweets during both interviews, we obtain approximately a sample of 14k tweets, 9k coming from US's interview, 5k coming from UK's interview. So, we have about 1 emoji every four tweets of the sample. Because of the different size of subsamples, *Fig6.1* does not want to show a numeric comparison between US and UK's interviews. Frequency on x axis of horizontal barcharts is reported in order to give an idea of the size of top emojis used.

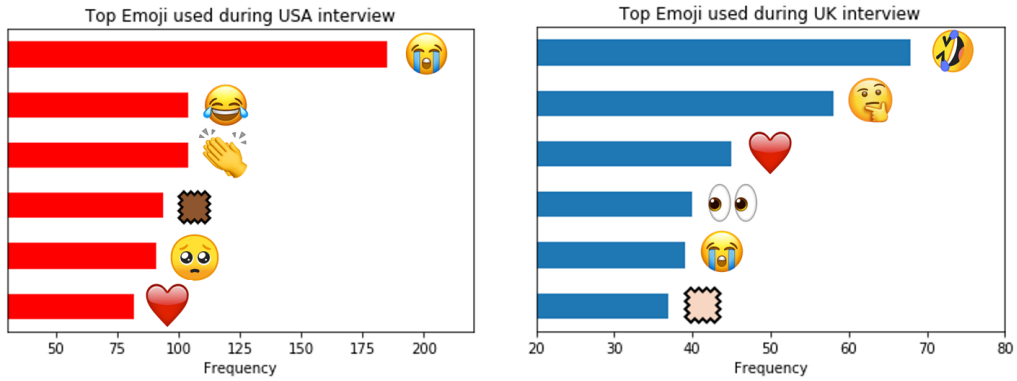

Fig6.1

Dominant emoji used during US's interview is sad face (another similar appears also in fifth position for USA and for UK), according to results obtained from emotion analysis shown previously, while during UK's interview the most used emoji is laughing emoji, which is definitely the most used emoticon in general by people (another similar appears in second position for USA). It could also have an ironic usage in relation with some topic encountered during the interview, maybe related to surprise emotion presented in emotion analysis. When you hear something totally different from your opinion, your response could be of surprise, that maybe could flow into a laugh.
The most interesting result coming from emoji analysis are brown shape in fourth position for USA and pink shape in sixth position for UK. This shape shows that during US's interview most of the hand emoji used were brown, while for UK's interview they were pink.

7 Retweets Analysis

In this final section of the paper we want to dedicate the right space to the *retweets*, which have been present in most of the analysis, and have played a leading role in some decisions we took. Moreover, as anticipated, retweets take about 40% of the collected sample, so they deserve an analysis. In this way will be shown retweets networks for both interviews, with the aim of understanding how users behave exploiting this particular functionality of Twitter, the retweet one. In Fig7.1 and in Fig7.2 are respectively represented US and UK's retweet networks.

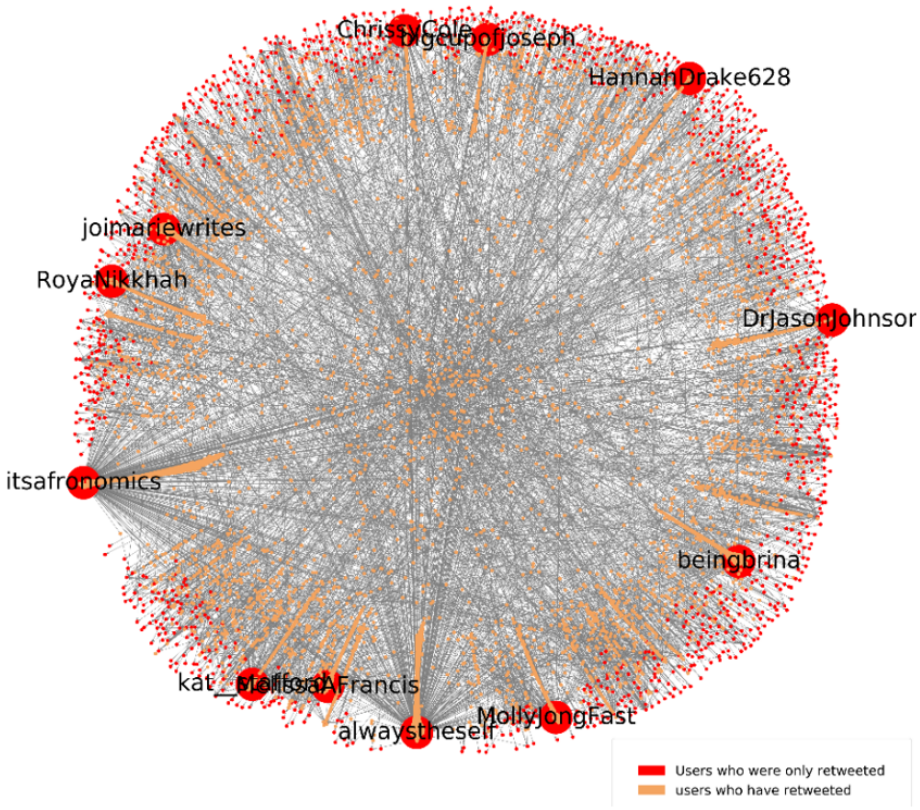

Fig7.1

As you can see, retweet networks are very similar. You can observe users who were only retweeted during interviews in a darker color, and users who have retweeted during the interviews in a lighter color. In the networks are

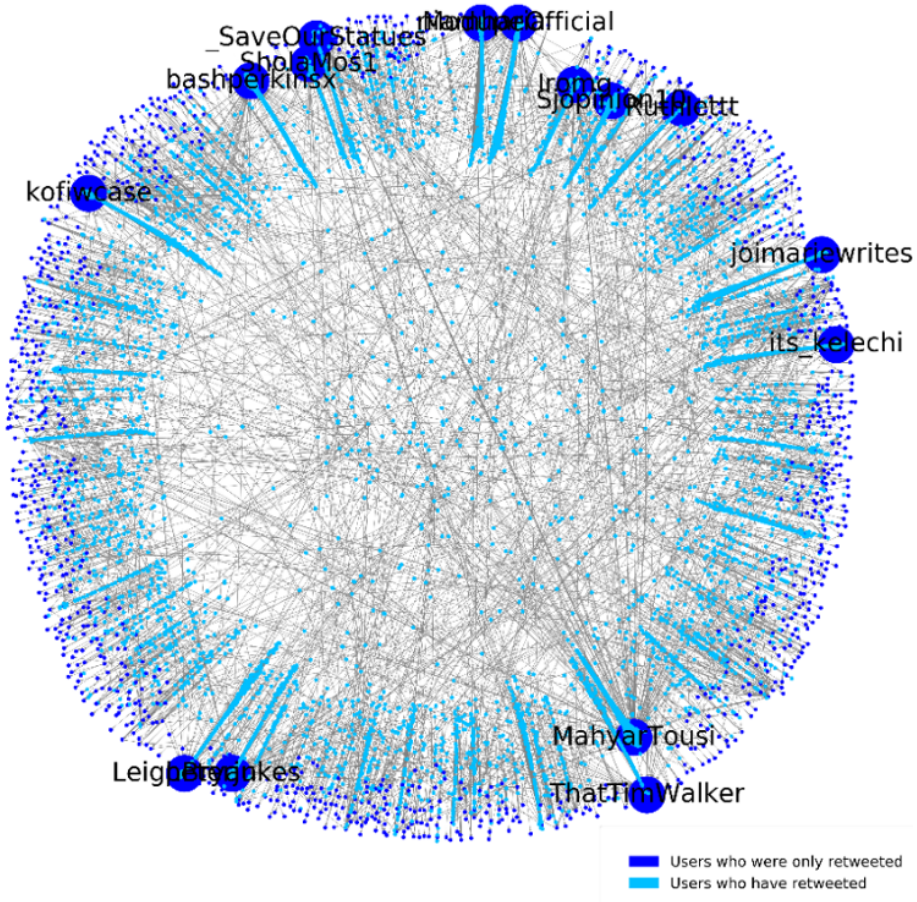

Fig7.2

put in evidence with their screen name users who have obtained more than 150 retweets during the interviews. We select this threshold to select *hubs* for the network. Users who have overcome this threshold could be considered as influencers.
Graphs are very disconnected, because of the presence of a few number of influencers who take most of the retweets, as you can see in Fig7.3.

Barchart's colors recall retweets network's colors related to US (hot colors) and to UK (cold colors). Analyzing barcharts, you notice very similar behaviour of the users in exploiting retweet functionality during both interviews. Most of the users obtain zero retweets during both interviews, and most of the users posted only one retweet during the interview it has

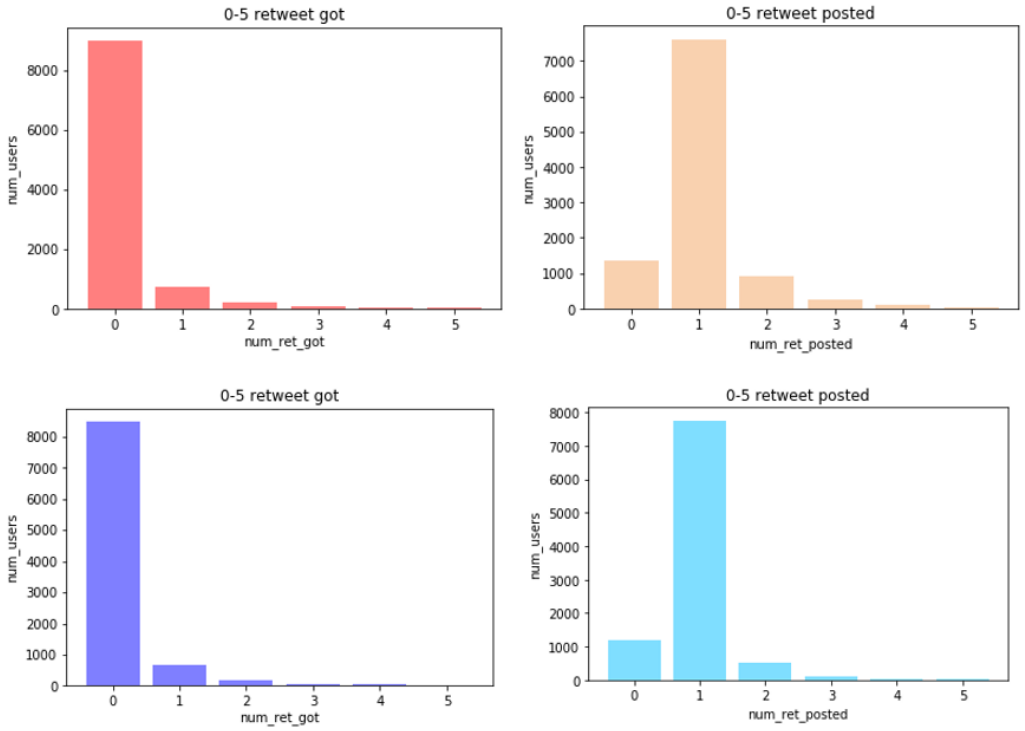

Fig7.3

whatched. This behaviour puts in evidence graphs's disconnection. In fact, most of the users tend to retweet influencers, who are very few in both networks.

8 Conclusions

In leading the analysis about Meghan and Harry's case has been interesting to provide answers to research questions we fixed. To sum up we have discovered that most followed interview has been the US one, and we have verified that tempers rise during interviews, particularly when big topics has been faced. We discovered that people emotions for the most part of both interviews belong to negative sphere of the emotions, with prevalence of sadness and anger referring to topics we analyzed. We have analyzed the importance of the retweets inside a tweets' collection. Retweets could give a lot of value to the analysis if treated in the right way.
Finding these answers has been possible to analyze how users behave during the interviews, not only from the point of view of feelings and emotions in relation to some topics encountered, but also from a technical point of view, through retweet's analysis for example. We know people in general tend to share what a person they like says, in some case they prefer sharing other's thought instead of sharing their own thought. In this analysis we have a confirmation of that situation, as we have seen in retweets network or focusing on racism topic. If on the one hand some people interfacing with Social Networks could be a little bit stopped in showing their real emotions in relations to relevant topics, on the other hand someone could also be facilitated in sharing his thought behind his keyboards. In this second sense Social Network could be a strong resource to discover and analyze users' attitudes.

References

[1] *Twitter API search, howpublished =* https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets.

[2] *Daily trends twitter 8 marzo, howpublished =* https://getdaytrends.com/it/italy/2021-01-17/23/.

[3] *Vader, GitHub Repository, howpublished =* https://github.com/cjhutto/vaderSentiment.

[4] *NRCLex documentation, howpublished =* https://pypi.org/project/NRCLex/.

[5] *Tableau Dashboard, howpublished =* https://public.tableau.com/profile/alessandro1080#!/vizhome/HMInterview_16210648383580/Dashboard1.