Social Media and Networks Analytics

Assignment 2

BREXIT

Table of Contents

Part1. Introduction

“Brexit” has been constantly in dispute both domestically and internationally since 2016 referendum. The term “Brexit” is composed of “British” and” Exit”, which refers to the withdrawal of the UK From the European Union (EU). The actual withdrawal has been postponed twice due to various issues. Although, UK is about to officially leave EU on 31 October 2019, there are still many issues not been settled yet. Also, studies have shown that social media have significant influence on political public opinions (Anstead & O'Loughlin 2015). Understanding public opinions is very important to untangle the Brexit problem.

The objective of this project is to analyse the public opinion on Brexit and its implication to the public using Twitter API. We are interested in find out who are the influential people, what do they say and why is it important. In this report, we will conduct descriptive analysis, sentiment, and social network analysis.

Part2. Methodology

Data Collection

To collect data from the twitter, we have used RESTFUL API to access **#Brexit** Tweets, as our main aim is to get the information about the users and their network i.e. their social connections, aimed to find out the most influential speaker about #Brexit on Social Media and we do not need live twitter streaming to do this analysis (Sandoval 2019).

To access data from twitter, we must set up twitter access by using python library "Tweepy" and pass the credential given by Twitter while creating developer account. Tweets can be searched from Twitter by using Cursor method which returns all the tweets containing #Brexit. The data is stored in the JSON file which is used in further processing. For ease of access, we have collected 1000 tweets on #Brexit.

Pre-processing and Data Cleaning

The tweets that we have collected is in raw form that contains special symbols like "@", "-", emojis, URLs, numbers and other unwanted patterns which creates problem while performing different Natural Language programming algorithms.

To do the data cleaning, we have incorporated the following steps:

- Convert text to lower case to avoid ambiguity
- Remove URLs from the text
- Remove Numbers
- Remove special characters and pattern from the text
- Remove words like https, RT, Via… etc.

The reason we need to perform these processing on our dataset is that we have many unwanted words that directly affects the further processing.

Tokenization is the processing of breaking the sentence into individual words called tokens which comes very handy while performing the Natural language processing tasks such as

Topic Modelling, sentiment Analysis etc. but tokens are just the segregation of sentence and give us the tokens of sentence as such.

To get the words that are present in NLP dictionary, we perform another step called Stemming which gives a one general word for several related words, for ex. Visuals, Visualizing, visualization etc. after stemming will result in Visual.

Sometime, but with stemming we get words that does not make sense, that happens specially with words like "does" which result in doe, to remove that problem, we performed the approach of Lemmatization, which gives lemmas that are actual words in English dictionary. For our analysis, we are using text that are available in "English" language only (Nlp.stanford.edu, 2019).

Since, the text from tweets contains words that are very frequent and appears almost in every document. Such words do not make any impact while performing text analysis.
We have a list of stop words available at NLTK corpus. Which we have used to remove these frequent words while performing tokenization and stemming so words like "is, am, are, the, RT, https:" do not appear in the corpus that we are using for analysis.
After performing the Data Cleaning and Pre-Processing, we get the text output as:
Now, the more predominant words are the ones that we are interested in not the stop words and special symbols.

Sentiment Analysis

Sentiment Analysis is a sub field of natural language processing that tries to extract and identify the opinion within a given text. The aim of sentiment analysis is to gauge the attitude, sentiments, evaluations, attitudes and emotions of a speaker/writer based on the computational treatment of subjectivity and polarity in a text.
Sentiment analysis is a hard task to perform mainly because of the fact that one sentence may contain multiple polarity and easier algorithms like TextBlob tends to neglect this scenario by negating the count of positive polarity with that of negative one and hence giving

a neutral polarity to sentence overall (Byrkjeland, M, Lichtenberg, F G D & Gambäck, B, 2018), (Spettel, S, & Vagianos, D, 2019).

To avoid this Problem, we have used a lexicon and rule-based sentiment analysis tool known as Valence Aware Dictionary and Sentiment Reasoner (VADER). Vader uses a combination of a sentiment lexicon which are generally labelled according to their semantic orientation as either positive or negative. VADER not only tells about positive or negative score, but it also gives information about how positive or negative a text is. It takes in consideration various subtleties such as punctuation marks and capitalisation.

• Topic Modelling

Topic Modelling is a type of statistical modelling for discovering the abstract "topics" that occurs in a collection of documents.To perform the Topic modelling for our twitter dataset, we have used Latent Dirichlet Allocation (LDA). It builds a topic per document model and words per topic model, modelled as Dirichlet distributions (Blei, D, Ng, A & Jordan, M, 2003).

The LDA's approach to topic modelling considers each document as a collection of topics in a certain proportion. Each topic as a collection of keywords, again, in a certain proportion.
Once we provide the algorithm with the number of topics, all it does it to rearrange the topics distribution within the documents and keywords distribution within the topics to obtain a good composition of topic-keywords distribution.
But how exactly a topic being extracted and represented?
A topic is nothing but a collection of dominant keywords that are typical representatives. Just by looking at the keywords, we can identify what the topic is all about.
The following are key factors to obtaining good segregation topics:

1. The quality of text processing.
2. The variety of topics the text talks about.
3. The choice of topic modelling algorithm.
4. The number of topics fed to the algorithm.
5. The algorithms tuning parameters.

- ### Social Network Analysis

For network analysis, we have done two network graphs to address different aspects of sociol network. Firstly, the interaction network graph is chosen since we want to find who are the influencers based on interaction (mention, reply, retweets). Determining the influencers for a topic simply based on followers count or number of likes is very likely to be biased, as those figures might not directly linked to the topic. Therefore, we consider a network graph based on interactions is more informative. The *Networkx* python package is used to model the graph. We recorded the user ids who posted the tweet, the users who replied, and the user ids that are mentioned in the tweets, then transformed them into graph nodes that are connected by edges. The graph captures how users interact based on the tweets related to Brexit. We focused on network degree centrality, which shows how important a node is within the network.

The graphs are viewed using spring layout which is a very popular and informative way to display graphs. The important nodes with maximum centrality can be distinctly visualised and discerned (Becker, M Y , Rojas, I , 2001).

Part3. Analysis & Discussion

Exploratory analysis

- ### Common words

From the common words (figure1) and word cloud (Figure 2), we can see some most frequently used words are closely related to our topic: Brexit. There is a name "johnson", which refers to Boris Johnson, former Prime Minister of UK. The Brexit referendum was held during his term and his name had been mentioned more than 200 times out of 1000 tweets. In addition, there are also some terms around Brexit deals, such as "proposal" and "deal". Also, some mention about "border" issue. We shall conduct topic modelling in the later part explore more topics related to Brexit.

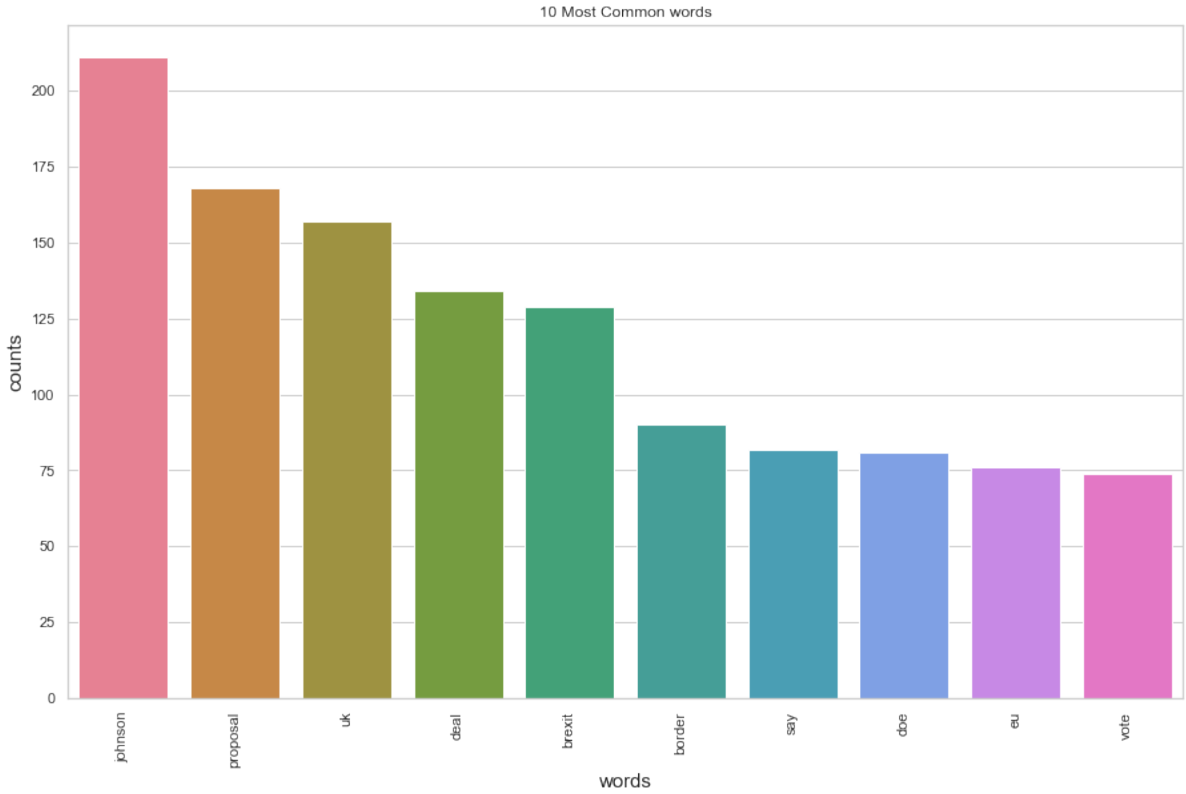


Figure 1 Top 10 common words

another custom
creates proposal
custom creates
friday nothing
resurrects reintroduces
one border
nothing risking
responding remain
border check
good worker
doe johnson
check johnson
right scrapping
vote responding
proposal resurrects
risking right
uk everyon
worker friday
brexit
proposal brexit
johnson jeopardi
deal
standard year
johnson uk
everyon vote
remain favour
scrapping agreement
jeopardi another

Figure 2 word cloud

- ## Frequent Hashtags

The list below displays the top 10 most frequently used hashtags in the data. Without doubt, "#brexit" is most frequently used. However, there are some other hashtags that shed some light on Brexit issue. First of all, "#scotland" and "#snp", Scotland is a country and part of the UK and SNP is the acronym for Scottish National party. According to the news articles, most of the Scottish are against Brexit (BBC NEWS 2016). It leads us to the fact that there are conflicts between UK and Scotland on Brexit. In addition, we noticed that there is a hashtag related to remaining in EU, such as "#remain". It is also worth discussing.

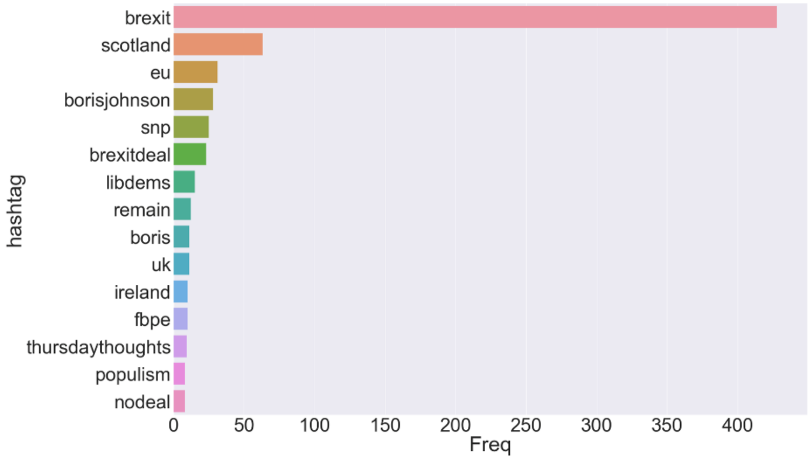


Figure 3

- ## Top 15 Frequently Mentioned User

Part of our project objectives is to find influencers on Brexit. Here, we visualized the top 15 frequently mentioned user in the data. Most of the users on the list (figure 4) are UK politicians. The one ranking at the top is screen name 'ianblackfordmp' which refers to Ian Blackford, a Member of Parliament (MP) and the SNP Parliamentary Leader at Westminster. There are also some news media that frequently been mentioned, such as RTE Politics and BBC Politics.

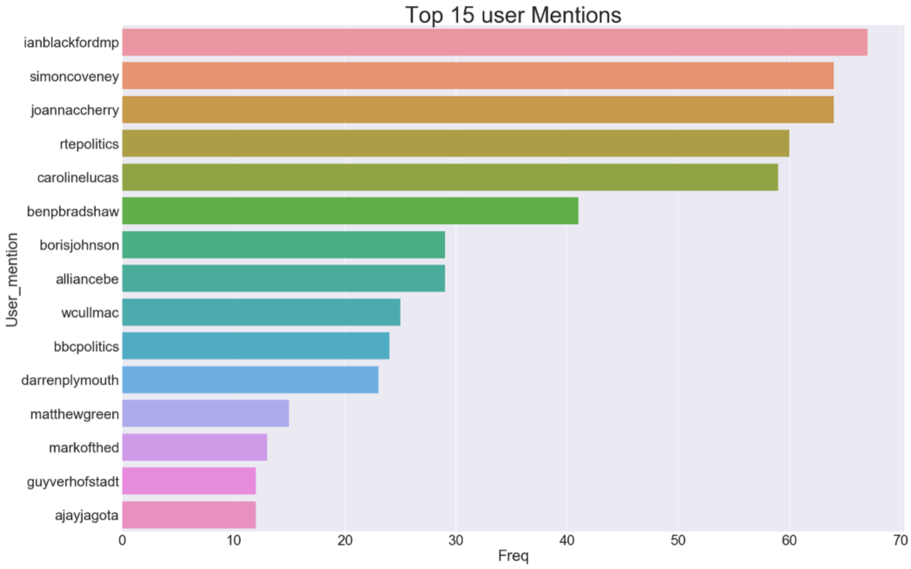
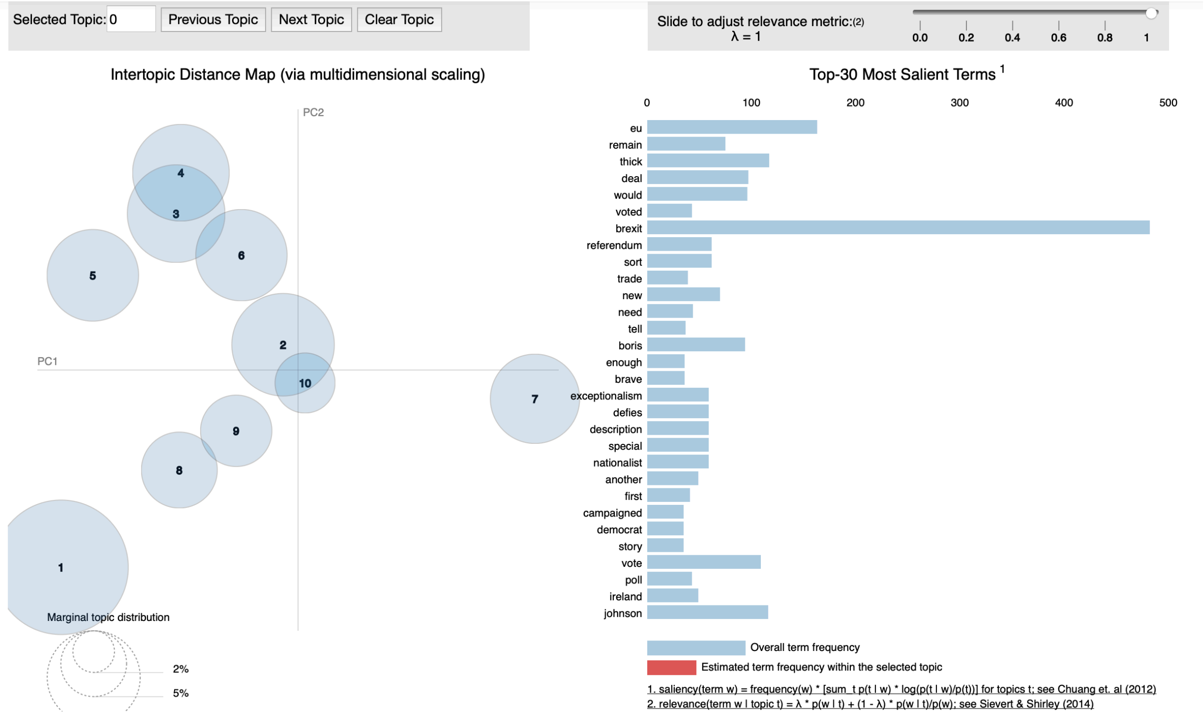


Figure 4 top 15 user mentions

- ## Topic Modelling

In our analysis, we have considered top 10 topics in the text and created an interactive pyLDAVis visualisation using LDA model which gives overall term frequency and estimated term frequency within the selected topic.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 5 PyLDAvis for topic modelling

Along with this model, we have also generated a word cloud to have a better look at the topics and terms that are most predominant.

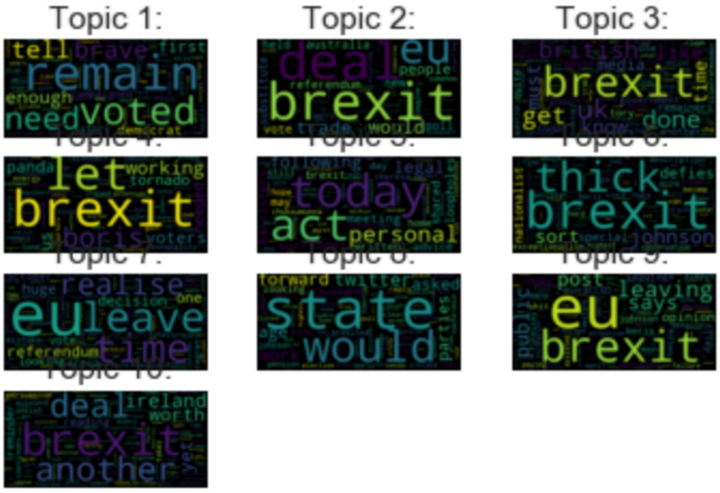


Figure 5 word cloud for topic modelling

As can be seen from the word graph, the most predominant term in the topics is “BREXIT”, which is very understandable as Brexit is the topic of discussion, but we have identified few words apart from Brexit such as EU, Remain, Voted, Deal, State, personal, leaving etc. giving us the idea about people opinions as from different people point of view, some people voted Britain to remain in EU while some Voted to leave. There were also some comments about Brexit being Boris Johnsons personal agenda.
Overall, these topics gives us the list of important terms that are more predominant during the discussions, or rather opinions about Brexit.

From Overall term frequency, The Words, Brexit, EU, remain, Johnson (Boris Johnson), Leave remain the most frequently appeared words.

- Sentiment analysis

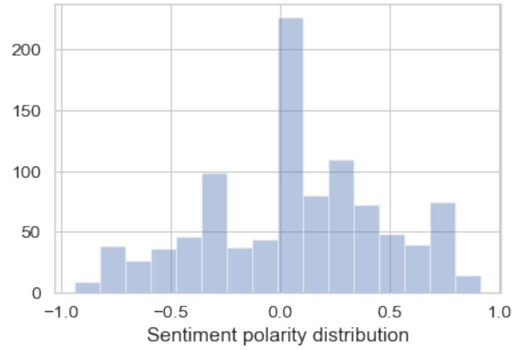
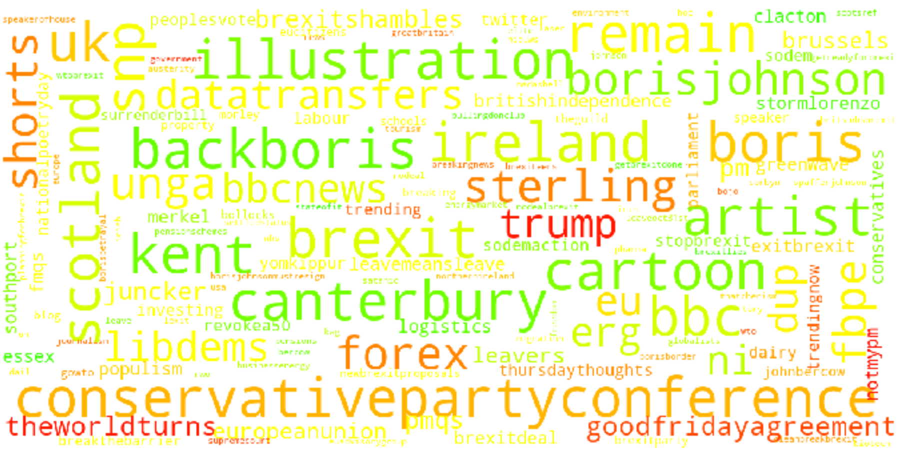


Figure 7 sentiment distribution

The polarity score is a float within the range [-1.0, 1.0], where -1 being negative end, 1 being positive end, and 0 being neutral. According to the distribution of the sentiment polarity, the histogram is widely spread out and has very obvious spike near 0. It shows the sentiment toward Brexit is mostly natural with some the opinions polarized across the spectrums.



WordCloud of hashtags based on sentiment polarity of containing tweets

Figure 8 word cloud of hashtags coloured by sentiments

The wordcloud of hashtag sentiment displays red color for negative sentiments, yellow for neutral and green for positive sentiments in the tweets containg the given hashtag. It can be clearly seen that the tweets with the hashtag trump are negative which implies that people don't like what trump says on the topic of Brexit. Other negative hashtags are related to the impact on the world and the political party which introduced Brexit. Overall though, Brexit has a rather neutral average sentiment associated with it.

- ## Social Network Analysis

1. Interaction Network Analysis

In the interaction network analysis, we will explore how people interact with each other on the Brexit topic. Essentially, we used three types of interactions in this part of the analysis, respectively "mention"," reply" and "retweet". The goal is to plot the graph of user interactions and find out who are the influencers in the network.

According to the summary statistics of the graph, there are 1105 nodes and 909 edges. The maximum degree of the graph is 63 and the minimum degree is 1. The nodes have average degree of 1.6. Although the graph is largely not connected, in the largest connected component, we have 367 nodes and 386 edges presented.

In order to find the central influencers in the network, we investigate the centrality of the graph. We find “Joanna Cherry QC MP” (yellow node in the graph) has a maximum degree centrality of 0.17 in this network. For both closeness centrality and betweenness centrality, the highest degree points to “Ben Bradshaw” (blue node), ” who is a British Labour Party politician.

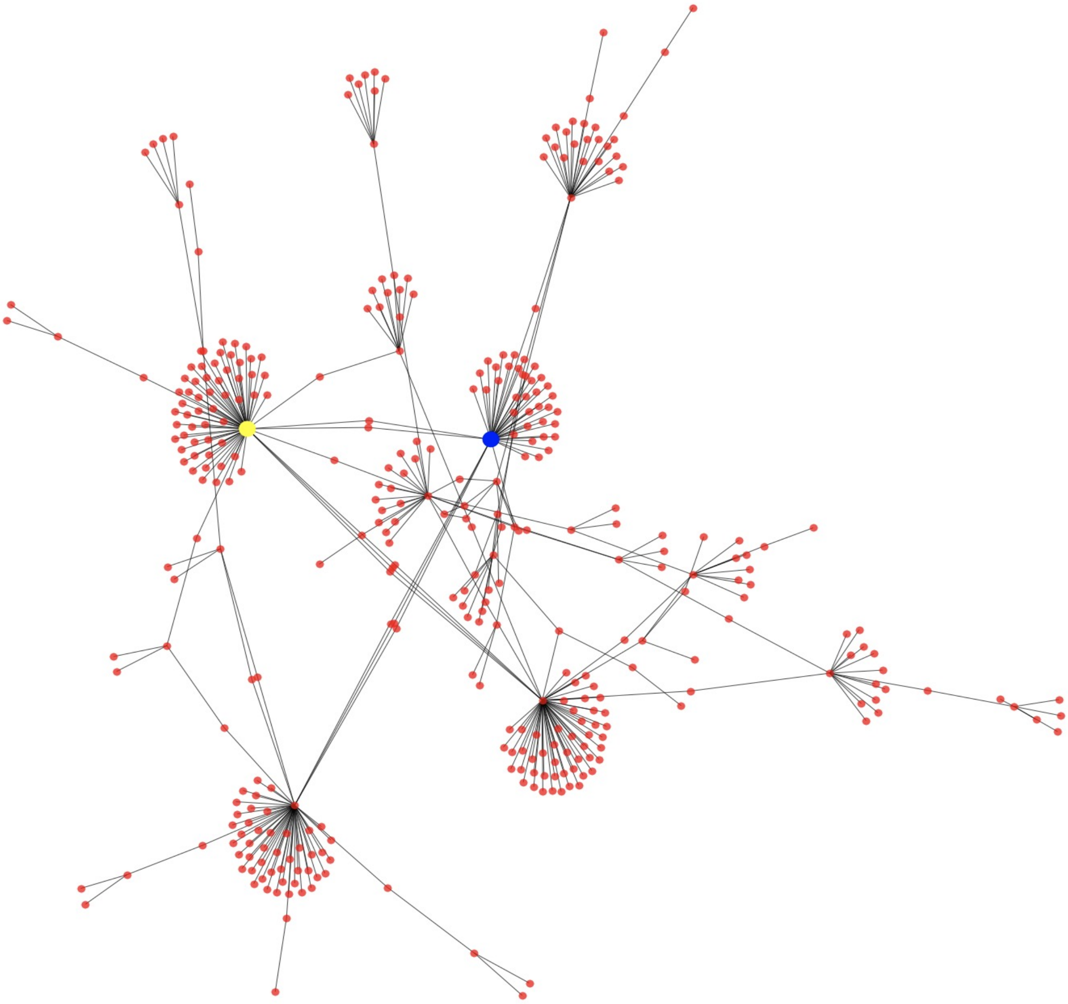


Figure 9 Interaction Network graph

2. Hashtag-User Network Analysis

The goal of the network graph of hashtags and users is to plot the graph of user interactions and find out the most popular and together occurring hashtags.

According to the summary statistics of the graph, there are 744 nodes and 897 edges. The maximum degree of the graph is 398 and the minimum degree is 1. Although the graph is largely not connected and there are 24 connected components in the graph, we have 678 nodes and 850 edges present in the largest connected subgraph.

We see a huge distribution around Brexit and breaking which appear together a lot. There are many people who use the hashtag Scotland but interestingly very few of them use hashtag Brexit. Further, a huge chunk of users also mentions the hashtag “eu” and “snp” together while not mentioning the hashtag Brexit.

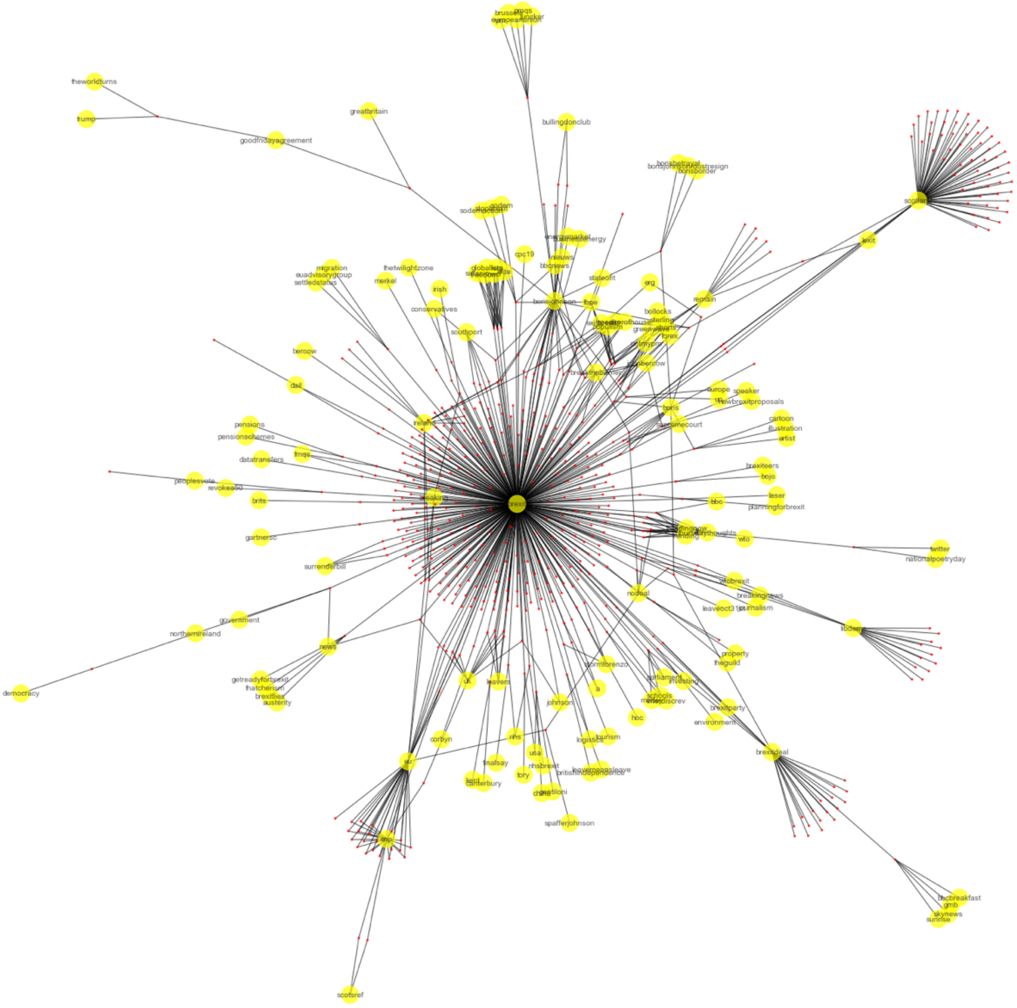


Figure 8 Hashtag-user network graph

Conclusions

Overall, Brexit is a very critical topic which concerns people from various backgrounds and interests. The tweets about this topic are usually people stating their opinion on the matter and there are almost equal number of people for and against the topic. Another topic that is commonly associated with Brexit is the vote of Scotland to separate itself from the United Kingdom. As for the influencers on Brexit topic, we identified “Joanna Cherry” as the top influencer based on maximum degree centrality in the network. She is SNP Spokesperson on Justice and Home Affairs and Queen's Counsel.

Reference

Anstead, N, & O'Loughlin, B 2015, ‘Social Media Analysis and Public Opinion: The 2010 UK General Election’. *Journal of Computer-Mediated Communication, 20*(2), 204-220.

BBC NEWS, 2016*, EU referendum: Scotland backs Remain as UK votes Leave*, BBC NEWS, viewed 12 October 2019,< https://www.bbc.com/news/uk-scotland-scotland-politics-36599102>

Sandoval, K, 2019, *REST vs Streaming APIs: How They Differ*, Nordic APIs, viewed 13 October 2019, <https://nordicapis.com/rest-vs-streaming-apis-how-they-differ/>

Nlp.stanford.edu, 2019, Stemming and lemmatization*,* viewed 13October 2019, <https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html>

Byrkjeland, M, Lichtenberg, F G D & Gambäck, B, 2018, ‘Ternary Twitter Sentiment Classification with Distant Supervision and Sentiment-Specific Word Embeddings’. *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis,* 97–106, Brussels, Belgium, October 31, 2018.©2018 Association for Computational Linguistics

Spettel, S, & Vagianos, D, 2019, ‘Twitter Analyzer—How to Use Semantic Analysis to Retrieve an Atmospheric Image around Political Topics in Twitter’, *Big Data Cogn. Comput. 2019,* 3(3), 38.

Becker, M Y, Rojas, I , 2001, A graph layout algorithm for drawing metabolic pathways , *Bioinformatics*, Volume 17, Issue 5, May 2001, 461–467

Blei, D, Ng, A & Jordan, M, 2003, Latent Dirichlet Allocation, *Journal of Machine Learning Research*, 3, Jan 2003, 993-1022