TEXT MINING

A CASE STUDY OF SENTIMENT ANALYSIS IN SOCIAL MEDIA NETWORK BREXIT VIDEO COMMENT

THE UNIVERSITY OF NOTTINGHAM NINGBO CHINA

ZEYUAN ZHANG

CONTENT

INTRODUCTION

What is sentiment analysis

Sentiment analysis, or sometimes being referred as opinion mining, is the method of using of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information (Bird at. el., 2009). Sentiment analysis has a board application scenario which can be widely applied to the area of commercial reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine.

The purpose of such analysis aims to determine the attitude of the speaker, writer, or any other subject with respect to some topic or the overall contextual polarity or emotional reaction to a document, interaction, or event. The sentimental score can also act as an index to predict future behavior of the subject being tested, therefore, its value could be maximized based on proper applications.

Why social media

With the widespread of the smart phone, social media assimilates into every aspect of people's life. It creates the possibility for everyone to concentrate and comment on social events. The rapid flow of the information does create a massive amount of data, in other words, they can be treated as precious treasure if under proper utilization. The prosperity of the data analysis company such as sprout social confirms this view (Nielsen and F. Å., 2011). To analysis the data generated by users with different background helps to make a more precise prediction, and this may benefit the commercial, especially for the more specific target market. Moreover, the information from social media network could also help to get a better understanding of the mechanism behind social events. However, the over exposure of the social media data may lead to unethical application and related regulation requires being established to protect the rights of citizens are not infringed.

Why Brexit

The reason brings the group members together is that the group shares the common interest in social media related to politics. As stated in the research abstract, the purpose is to analysis people's reaction towards certain social events and visualize the outcome. The sentimental analysis is used as a helpful tool in this condition, what the research group would like to achieve is the comparison of the sentimental analysis with other index from economic and political areas to illustrate the overall influence of a certain social event after Brexit. Under such consideration, the analysis topic of Brexit would be an ideal choice.

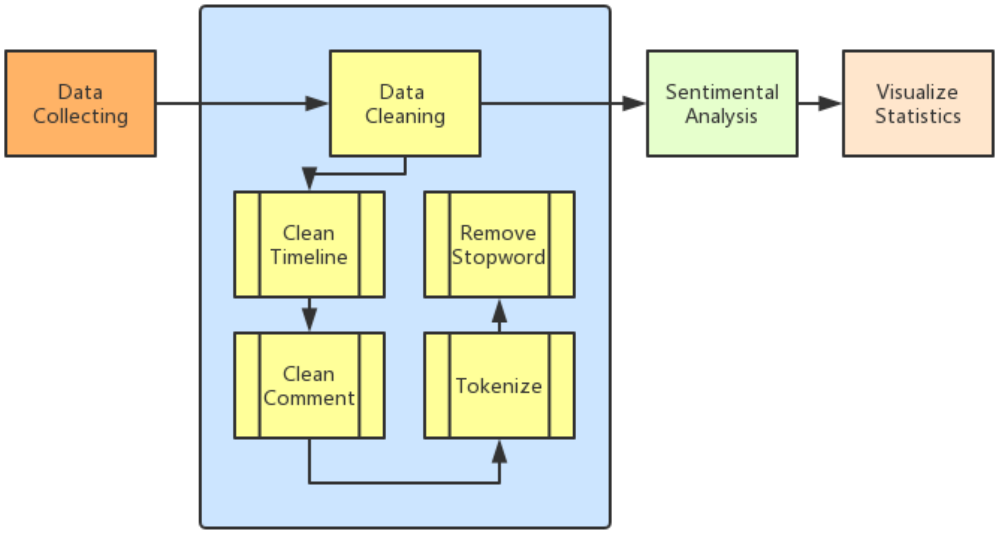
What problems of the related topic

In the initial plans, Twitter post with specific hashtag like #Brexit would be the first choice. However, based on the consideration of protecting user personal information and other related issues, the official Twitter API only allows to view tweets within the last one week. For historical data, there are other data companies provide such commercial service with an unacceptable price. Therefore, the YouTube video comment become the one last choice (YouTube, 2016). The related issue of YouTube comment could be concluded as follows:

1. As social media, all the information has timeliness property, just as the former assumption, most of the comment concentrate on the first month after it was released.
2. Limited by the legal condition and storage mechanism from YouTube, it is impossible to locate the exact date of each comment in timeline, only to month.

METHOD

Preliminary preprocessing of the data is described as follow:



Pic 1. Working flow. *Plotted by author*

Data collecting

The corpus used un this exploratory study is the comment from the Brexit video on YouTube under the API provided by YouTube. Therefore, there should be no copyright issue to be considered in this condition. The corpus is saved as json format for sentimental analysis and further application.

Data cleaning

The timeline dragged from the json data is in the form like: 2017-07-29 09:45:07.908000. For the sake of using the timeline as index, the timeline form should be simplified to the format of 2017-07-29. And in this case, dt.normalize( ) function should be applied.

Next step is to clean comment text. The rule behind is to save only the strings formed by a-z and A-Z, delete any other signs and leave only one space between two neighbor words. In this case, re.sub function should be applied. It returns the string obtained by replacing the leftmost non-overlapping occurrences of pattern in string by the replacement repl. If the pattern isn't found, string is returned unchanged.

Based on the clean text step, it saves a lot effort for doing tokenization. Create a loop to split and lower case each word and return to a token list. Set up a new column under the name of tokens.

Further step is to import nltk to remove the stop word in tokens. Stop words refers to the words has no meaning in sentimental analysis. Purify tokens helps to reduce the useless work when doing emotion analysis and therefore increase the efficiency.

Sentimental analysis

Import panda to read the csv file containing the sentimental score of the words in the dictionary. Again, create a loop to calculate the sentimental score of each comment and export the result to a list. Set up a column in panda data frame named sentiment.

Visualization

In the visualization step, import seaborn.heatmap function to represent how sentimental scores changes with the estimated period (Seaborn, 2017).

RESULTS

Popularity of terms

After cleaning the timeline, the comment can be sort out by time. In this case, the popularity of terms is ranked by numbers of replies under the video. Generally, more replies represent higher popularity. It makes sense that majority of the replies concentrate on the first month after the Brexit video being uploaded. After which, it decreases dramatically for people begin to lose interest and they change the attention to other more practical issues related with Brexit. However, during several periods, it can be easily distinguished that both google search trend and the comment of the YouTube video have experienced an increase. A pre-assume can be made here that the increase

in both video comment and google search trend are induced because of certain political events. There are two major varieties in the timeline. Statistical analysis proves this point in the following aspects:

1. The first of period when comments experience fluctuation is in Nov 2016, from both google trend and YouTube comment have seen a significant peak time. The other one is the period of the time start from the late April to early May 2017. From both sides, after the first two month, Brexit the topic seems to cool down somehow. However, in these two period it draws back public attention in some extent.

2. The changes of average sentimental score in the same two periods of time. From the October to November 2016, the sentimental score declines dramatically from positive 0.3 to negative 1.6 with comment tendency from natural slightly positive to obvious negative. This means a dramatical shift of public opinion. In the first month of 2017, the negative sentimental comment reaches its peak, up to 2.7 negative. Similarly, another dramatical change in emotion happens in between March which change back from normal negative 0.9 to very strong positive 2.0

3. From the opinion polling from YouGov, the time late November 2016 when public opinion of Brexit moves from support to disagreement for the first time. For the period mid-January 2017, the pro-Brexit reaches its highest point where 13 percent of the interviewee holds positive altitude. Same changes happen in the time late April to mid-May.

All these observation points to the same period when public opinion experience fluctuation. Next, analysis will be applied to point out the influence of political event with public opinion towards Brexit.

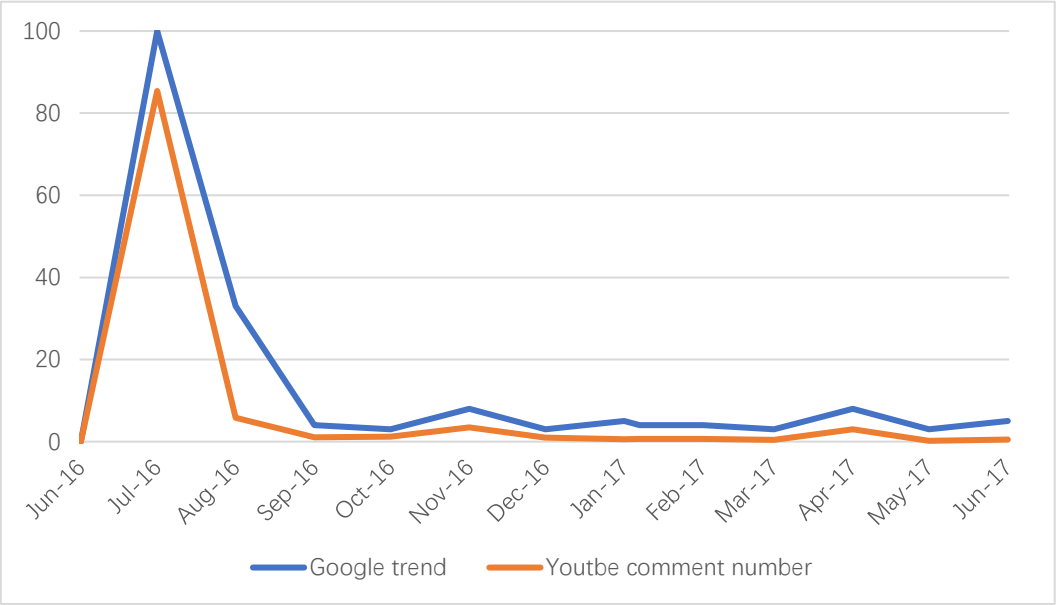
Post-referendum opinion polling [edit]

Following the EU referendum, there have been several opinion polls on the question of whether **the UK was "right" or "wrong" to vote to leave the EU**. The results of these polls are shown in the table below.

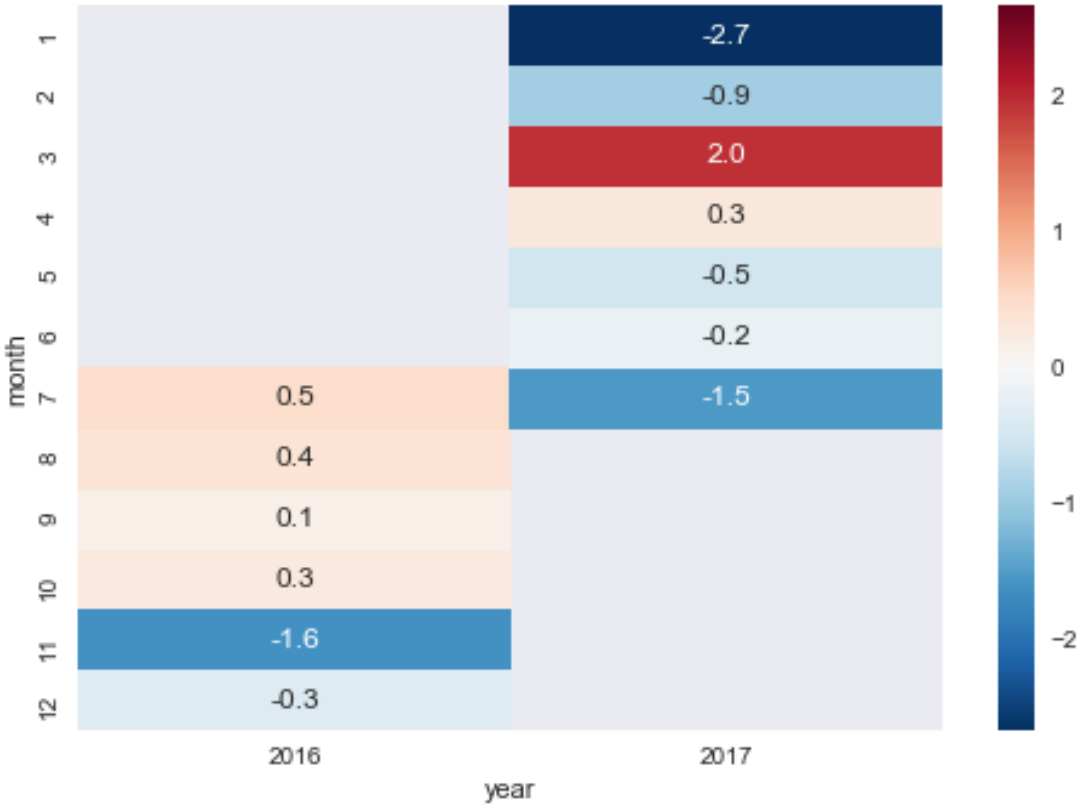
Date(s) conducted	Right	Wrong	Undecided	Lead	Sample	Conducted by	Polling type	Notes
V • T • E								
13 Jul 2016	***Theresa May becomes Prime Minister of the United Kingdom.***[207]							
1-2 Aug 2016	**46%**	42%	12%	4%	1,722	YouGov	Online	
8-9 Aug 2016	**45%**	44%	12%	1%	1,692	YouGov	Online	
16-17 Aug 2016	**46%**	43%	11%	3%	1,677	YouGov	Online	
22-23 Aug 2016	**45%**	43%	12%	2%	1,660	YouGov	Online	
30-31 Aug 2016	**47%**	44%	9%	3%	1,687	YouGov	Online	
13-14 Sep 2016	**46%**	44%	10%	2%	1,732	YouGov	Online	
2 Oct 2016	***Theresa May makes Conservative Party Conference speech, announcing her intention to invoke Article 50 by 31 March 2017.***[97]							
11-12 Oct 2016	**45%**	44%	11%	1%	1,669	YouGov	Online	
19-20 Oct 2016	**45%**	44%	11%	1%	1,608	YouGov	Online	
14-15 Nov 2016	**46%**	43%	11%	3%	1,717	YouGov	Online	
28-29 Nov 2016	44%	**45%**	11%	1%	1,624	YouGov	Online	
4-5 Dec 2016	**44%**	42%	14%	2%	1,667	YouGov	Online	
18-19 Dec 2016	**44%**	**44%**	12%	0%	1,595	YouGov	Online	
3-4 Jan 2017	**45%**	44%	11%	1%	1,740	YouGov	Online	
9-10 Jan 2017	**46%**	42%	12%	4%	1,660	YouGov	Online	
9-12 Jan 2017	**52%**	39%	9%	13%	2,005	Opinium	Online	
17 Jan 2017	***Theresa May makes Lancaster House speech, setting out the UK Government's negotiating priorities.***[206]							
17-18 Jan 2017	**46%**	42%	12%	4%	1,654	YouGov	Online	
30-31 Jan 2017	**45%**	42%	12%	3%	1,705	YouGov	Online	
12-13 Feb 2017	**46%**	42%	12%	4%	2,052	YouGov	Online	
21-22 Feb 2017	**45%**	**45%**	10%	0%	2,060	YouGov	Online	
27-28 Feb 2017	**45%**	44%	11%	1%	1,666	YouGov	Online	
13-14 Mar 2017	**44%**	42%	15%	2%	1,631	YouGov	Online	
10-14 Mar 2017	**49%**	41%	10%	8%	2,003	Opinium	Online	
1-15 Mar 2017	**46%**	41%	13%	5%	1,938	GfK	Online	
20-21 Mar 2017	**44%**	**44%**	12%	0%	1,627	YouGov	Online	
26-27 Mar 2017	**44%**	43%	13%	1%	1,957	YouGov	Online	
29 Mar 2017	***The United Kingdom invokes Article 50.***[205]							
5-6 Apr 2017	**46%**	42%	11%	4%	1,651	YouGov	Online	
12-13 Apr 2017	**45%**	43%	12%	2%	2,069	YouGov	Online	
18-19 Apr 2017	**46%**	43%	11%	3%	1,727	YouGov	Online	
20-21 Apr 2017	**44%**	**44%**	12%	0%	1,590	YouGov	Online	
25-26 Apr 2017	43%	**45%**	12%	2%	1,590	YouGov	Online	
2-3 May 2017	**46%**	43%	11%	3%	2,066	YouGov	Online	
9-10 May 2017	44%	**45%**	11%	1%	1,651	YouGov	Online	
3-14 May 2017	**45%**	41%	14%	4%	1,952	GfK	Online	
16-17 May 2017	**46%**	43%	11%	3%	1,861	YouGov	Online	
24-25 May 2017	**46%**	43%	11%	3%	2,052	YouGov	Online	
30-31 May 2017	44%	**45%**	11%	1%	1,875	YouGov	Online	
5-7 Jun 2017	**45%**	**45%**	10%	0%	2,130	YouGov	Online	
8 Jun 2017	***United Kingdom general election, 2017***							
12-13 Jun 2017	44%	**45%**	11%	1%	1,651	YouGov	Online	
19 Jun 2017	***Brexit negotiations begin.***[204]							
21-22 Jun 2017	44%	**45%**	11%	1%	1,670	YouGov	Online	
10-11 Jul 2017	**45%**	43%	12%	2%	1,700	YouGov	Online	
18-19 Jul 2017	**43%**	**43%**	14%	0%	1,593	YouGov	Online	
31 Jul-1 Aug 2017	**45%**	**45%**	10%	0%	1,665	YouGov	Online	

Pic 2. Post-referendum opinion polling

Source: http://www.consilium.europa.eu/en/policies/eu-uk-after-referendum/



Pic 3. Google trend and YouTube comment popularity. *Plotted by author*



Pic 4. Sentiment analysis score *Plotted by author*

Analysis

Time 1: November to December 2016

In November 2016, Prime Minister Theresa May proposed that Britain and the other EU countries mutually guarantee the residency rights of the 3.3 million EU immigrants in Britain and those of the 1.2 million British citizens living on the Continent, in order to exclude their fates being bargained during Brexit negotiations. Despite initial approval from a majority of EU states, May's proposal was blocked by European Council President Tusk and German Chancellor Merkel (European Council, 2016).

This can be proved by statistical data. Firstly, the rebound of public attention implies public reaction. This kind of “blocked action” kind of annoy the Britain people and directly lead to negative comment on YouTube video to express the dissatisfied mood for the unfriendly reaction from the European continent. The post-referendum opinion polling, from the first time, against the behavior to leave EU. Public worries of the potential opposition probably the factors behind- the increasingly criticize from the other side weakens the public confidence for the recovery of British depression economy. Besides, during the past month position from EU, Brexit seems to conquer with more barriers than expected. The tough talk that from the European continent again induce public uncertainty towards the outlook of Brexit and this is reflected in the opinion polling.

Time 2: April to May 2017

On 29 April 2017, immediately after the first round of French presidential elections, the EU27 heads of state accepted, without discussion, negotiating guidelines prepared by the President of the European Council. The guidelines take the view that Article 50 permits a two-phased negotiation, whereby the UK first needs to agree to a financial commitment and to lifelong benefits for EU citizens in Britain, before the EU27 will entertain negotiations on a future relationship. In the requested first phase of the withdrawal negotiation, the EU27 negotiators demand the UK pay a "divorce bill", initially estimated as amounting up to £52bn and then, after additional financial demands from Germany, France, and Poland, amounting to £92bn. Nevertheless, a report of the European Union Committee of the House of Lords published on 4 March 2017 states that if there is no post-Brexit deal at the end of the two-year negotiating period, the UK could withdraw without payment. Similarly, the Prime Minister insisted to EU Commission President Juncker that talks about the future UK-EU relationship should start early and that Britain did not owe any money to the EU under the current treaties.

This can be proved by statistical data as well. The guideline from EU side again brings a lot argument in both sides. From the UK sides, the request for a “divorce fee” sounds like a blackmail action. This huge compensation, the retaliation will damage the future relationship between EU-UK. Furthermore, it induces the more public untrusty of each side and simulate the pessimism of the market which can be shown from the sentimental

analysis score changing from positive to negative. From the EU sides, the retaliatory request for kind of separation fee in necessary to deter the further potential separation of EU inside.

Weakness and improvement

This assignment is based on the panda data frame of the Brexit YouTube video comment and the aim is to analysis the sentiment behind and how its related to the social events after Brexit. Therefore, the focus is on the outcome, the score of sentimental analysis. However, two problems are closely related with such analysis.

1. Sentimental score can reflect the altitude of commenters but not the altitude of whether support of against Brexit. For instance, comment can be like saying stupid EU to apply too much additional condition which is no other than rip off behavior or foolish UK people to vote for leaving EU. Both comments could be ranked as negative sentimental score whereas with definitely opposite altitude towards Brexit. Based on the observation of the comment, the sentimental score does have a close relationship with the political attendance, but it really depends on political events. During the period of November 2016, January and April 2017, negative sentimental score represents the tendency for against Brexit and positive for approving. While in other situations, it is the opposite. To conclude, further explore the relationship between sentimental score and public altitude depends on the political event behind and there does not exist a general standard.

2. Another data could help to fully analysis the panda data frame such as the number of replies and how many upper words written in the comment. Usually, in social media and any other places, Sentence written in capital letters is a symbol of emphasize sentiment and express for strong argument. To count for such usage of capital letter indeed provide more accuracy of the analysis but require extra work. Based on the observation, the comment with strong sentimental scores, no matter how positive or negative the score are, does use a lot capital letters in the comment. Whereas the natural comment tends to use less such emphasize. Based on this view, it is not that need to count for capital letters and increase or decrease the sentimental score. Otherwise, such procedure will further cause the separation of the sentimental analysis score.

Conclusion

During the analysis of social media, specifically under the topic of Brexit. It shows how promising that text mining, the sentimental analysis could be applied to analysis the public opinion towards certain social hot topics. From the above analysis clearly shows the relationship between sentimental respond and public opinion of Brexit. Indeed, there exists certain weakness and could be improved further. However, aside from opinion

polling, collecting data from social media to fully analysis the target group process great potential to discover. With the awareness of citizens privacy, related laws should be introduced to restrict the abuse of data and in return protect public.

From the point view of Brexit, such black events which could never be able to happen in the past become reality nowadays. Start from president election in the United State, Brexit in the UK, then the arise and boom of extreme right wing in the European continent. People seems to make unbelievable decision again and again? But the truth is that they are making the 'right choice', the right choice for themselves. During the past decade in the new century, as the global citizens, we enjoy the advantage from the globalization, but what about the disadvantages? This irreversible process benefits the monopolist, the affluent class, maybe in a great degree of the middle class, but what about the others, the common people are left behind? Take the refugee example in Europe, large promotion of the tax is being implemented on this issue whereas it should be invested in other public service to benefit its own citizens. When people begin to question these issues, disagreement appears. The real thought from public are not truly reflected on newspapers, on the traditional news agency which being described as fake news by Trump, but the real thought could be obtained from the social media, from the tweet, from the comment that people left on every corner in the internet. This is the future of the text mining, read public minds to help the nation to understand how people think of their policies, what is the influence of them and who can we avoid such black swans become the norm in our society. By seeking the help of text mining, a prosperous future could be expected.

Open Github tmgu17_SIS at: https://github.com/zeteschang/tmgu17_SIS

Reference:

Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python (1 edition). Beijing. Cambridge Mass. O'Reilly Media.

Brexit the movie full film. Available at: https://www.YouTube.com/watch?v=UTMxfAkxfQ0 (Accessed date: 3rd Aug 2017).

European Council (2016). Council of the European. UnionEU-UK after referendum. Available at: http://www.consilium.europa.eu/en/policies/eu-uk-after-referendum/ (Accessed date: 6th Aug 2017).

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. arXiv preprint arXiv:1103.2903.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2(1-2), 1-135.

Sentiment analysis for YouTube channels with NLTK. Available at: https://datanice.wordpress.com/2015/09/09/sentiment-analysis-for-YouTube-channels-with-nltk/ (Accessed date: 7th Aug 2017).

Sweigart, A. (2015). Automate the Boring Stuff with Python: Practical Programming for Total Beginners (1 edition). San Francisco: No Starch Press.

Tutorial for seaborn. Available at: http://seaborn.pydata.org/generated/seaborn.heatmap.html (Accessed date: 8th Aug 2017).