Analyzing the national variations in the political sentiments using Twitter data – A case of Scottish Independence

Author: Monsuru Adepeju

Introduction

In the three and half years since the Brexit referendum, the conversations around Scotland’s independence, dampened momentarily by a failed referendum in 2014, has been reinvigorated with the ‘certainty’ of Brexit, thanks to the December 2019 UK general election. Scottish independence is the political movement for Scotland to become a sovereign state, independent from the United Kingdom. Following the general election, Twitter has become a major platform where conversations regarding the potentials of another Scottish referendum take place. These conversations (posts & tweets) can be identified based on the hashtags, such as ‘#Indyref2’ and ‘#scottishreferendum. In this article, I downloaded and used this dataset (having excluded ‘retweets’ and ‘replies’) between the January 1 and January 25 2020 (days before Brexit) in order to examine and compare the sentiments, across the four constituents nations (i.e. England, Wales, Northern Ireland and Scotland) of the UK.

While there has been a considerable amount of studies on how sentiments are expressed in conventional media, such as new articles, online reviews, and blogs, how sentiments or feelings are communicated in casual, length-restricted microblogging or social media, such as Twitter has been relatively less studied. Twitter provides a valuable source of diverse range of political sentiments subject on political subjects. With the development of programmable sentiment dictionaries (\*) through which a sentence can be analyzed to derive the underlying sentiments, there now opens new potentials to derive novel ideas from this new form of communication. This paper makes use of two tools in order to derive new knowledge regarding Scottish independence from Twitter data. They are (i) Word clouds and (ii) R-Sentiment classifier. Word clouds has been used to fill in as a beginning stage for a more profound analysis [1-3]. They have been used primarily to highlight the relative importance of words used in communication (\*). One of limitation of word clouds is that they give a simplistic rundown of disengaged words without considering the context of usage or the relationships between words. However, I think there is a bigger potential to this static yet apt dataview in numerous analysis contexts.

Leveraging the availability of programmable sentiment lexicons, I developed an R-sentiment classifier which categorises to be scored according to the sentiment expressed in it. Sentiment analysis is a developing area of Natural Language Processing with research extending from document level characterization [4] to taking in the extremity of words and phrases [5, 6].

In this work, I utilize both word clouds and sentiment lexicons in order to. In this paper, we analyze one such prevalent microblog or social media called twitter and build R models for characterizing "tweets" into positive, negative and unbiased sentiment and also create word cloud to find out the most frequently used term. For twitter sentiment, we assemble models for twitter authentication, and then we will pull the data from twitter. Here we will use a political figure to analyze sentiment what type of words are being used by him in everyday life to figure out actually what is happening in his mind. By using the R models, we will basically create a graph of positive, negative and neutral words used by the twitter user. To generate word cloud, we will first use R model to authenticate twitter. Then we will pull twitter data of a famous phone company. Then we will process the twitter data in a way that we can create a word cloud based on the dataset. The finalized word cloud will picture what the company is actually thinking. Meaning which words are being used frequently on this particular twitter account.

2.2 Political Sentiment Analysis

On-demand interest in the online political sentiment analysis is to predict the result of the election. Tumasjan et al. focused in his research on the German federal election in 2009 [14]. Their investigation identified that the Twitter became as a platform for analyzing sentiment and also predicting the outcome of the election [14]. They examined on around one lakh political tweets identifying either a politician or a political party. They used LIWC2007 tool for extracting sentiment from the tweets [14, 15]. LIWC is accurate text analysis software developed to reveal people thoughts, emotions, cognitive and personality using text samples [15]. They concluded that the number of tweets is directly proportional to the chances of winning the election. O’Connor investigated on the people’s remark measured from polls with opinion measured from microblogging sites [16]. They analyzed various political sentiment from 2008 - 2009, and result in the correlation between the frequencies of sentiment word in simultaneous with the twitter tweets. Choy et al. proposed an application of online sentiment analysis to predict the vote percentage for each of the candidates in the Singapore presidential election of 2011 [17]. Wang et al. presented a real-time sentiment application system for U.S. presidential election of 2012 based on political tweets extracted from Twitter [18]. Ringsquandl and Petković worked on the campaign topics of presidential candidates of the Republican Party in the USA. Their research, they introduced the amalgamation of the frequencies of noun phrases and their PMI measure with a constraint on aspect extraction and concluded that the semantic relationship between politicians and their topics holds. Their result has also improved the accuracy of the aspect extraction [19]. Elghazaly, Mahmoud, and Hefny also focused on Egypt Presidential Elections ‘12 based on Arabic text classification based on WEKA application. They concluded that the Naïve Bayesian method gives the highest accuracy with the lowest error rate [20]. In general state election in 2016, Sharma and Moh researched on Hindi Twitter for predicting of Indian state election. They have used 42,235 tweets using Twitter Archiver tool to extract tweets in Hindi language and examined by Support Vector Machine (SVM), Dictionary Based and Naive Bayes algorithm which classified into three categories like “positive”, or “negative” or “neutral” [21]. The objective of our research is to analyze peoples’ sentiment for determining the results of assembly election of Gujarat in 2017 using deep learning method tools available as web service by ParallelDots Inc [22].

3.2 Data Processing

Before applying any of the sentiment or emotion analyzing method [25, 26], we perform data preprocessing. Data preprocessing is a major step in text mining that produces a higher quality of text classification and reducing computational complexity [27]. The following points have mentioned as data preprocessing steps of the tweet.  Remove all hashtags (e.g. #topic), screen name (e.g. @username) and all URLs (e.g. www.xyz.com)

 Remove all punctuations, symbols, and numbers

 Replace any non UTF-8 by space

 Remove Stopwords

 Substitute all the emoticons with their sentiment

 Change text to lowercase

 Replace words with their stems or roots

4. Experimental Setup and Result Analysis

We performed the Emotion Analysis using syuzhet in CRAN package which is based on

NRC Emotion Lexicon [28]. Figure 4 and figure 5 show the distribution of emotions of Vijay

Rupani and Bharatsinh Madhavsinh Solanki during the Gujarat Legislative Assembly

election, 2017.

In figure 4, more than 300 tweets express positive sentiment and around 160 tweets indicate

negative sentiment whereas in figure 5, more than 100 tweets express positive and around 70

tweets represent negative sentiment. In figure 4, more than 200 tweets express trust whereas

in figure 5 around 90 tweets express trust as an emotion. As can be seen, the disgust emotion

in figure 4 is second least emotion score compare to the figure 5. Similarly, joy emotion is

placed better in figure 4 than the emotion score in figure 5.

Collecting, Storing and Processing data

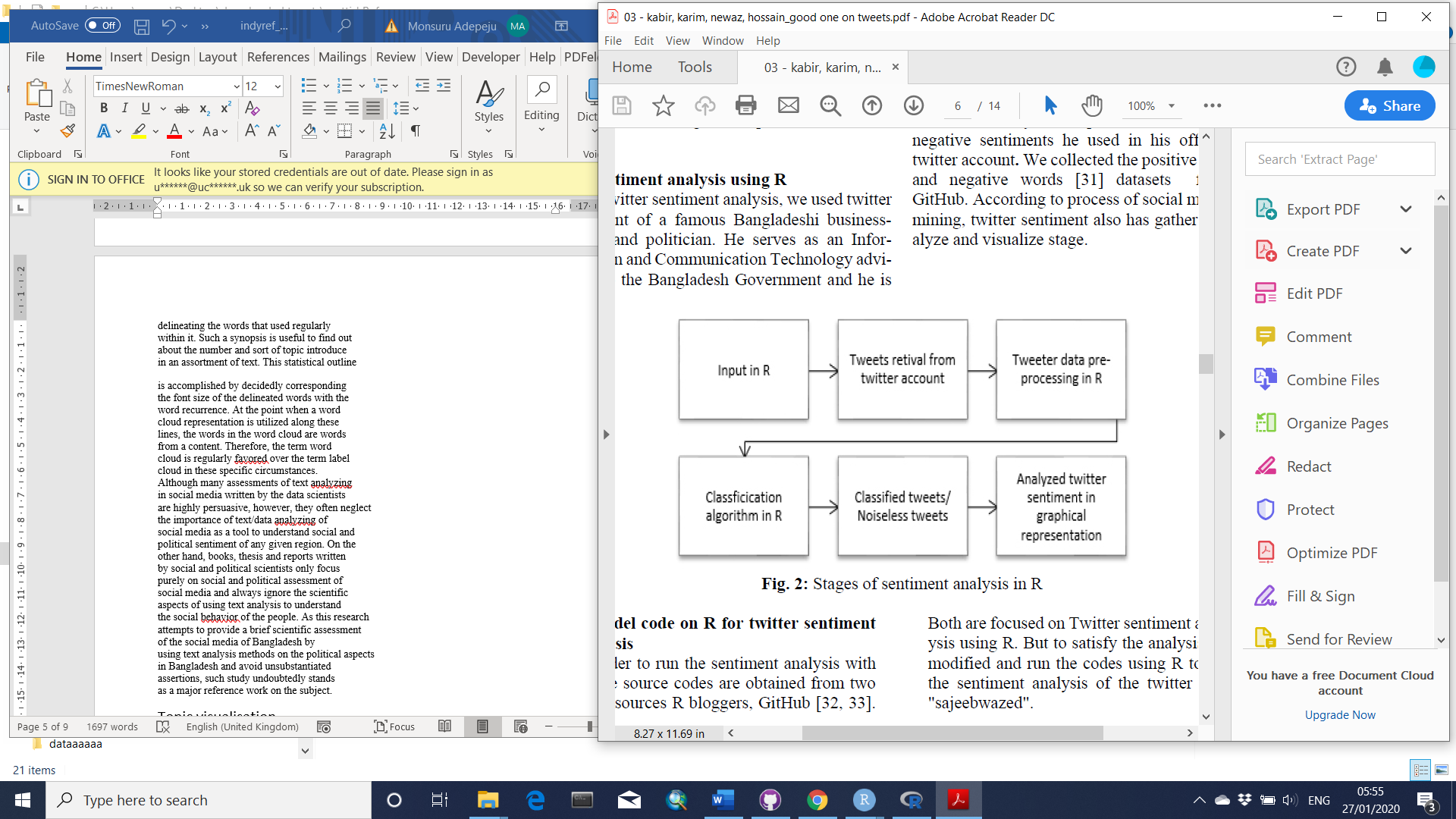
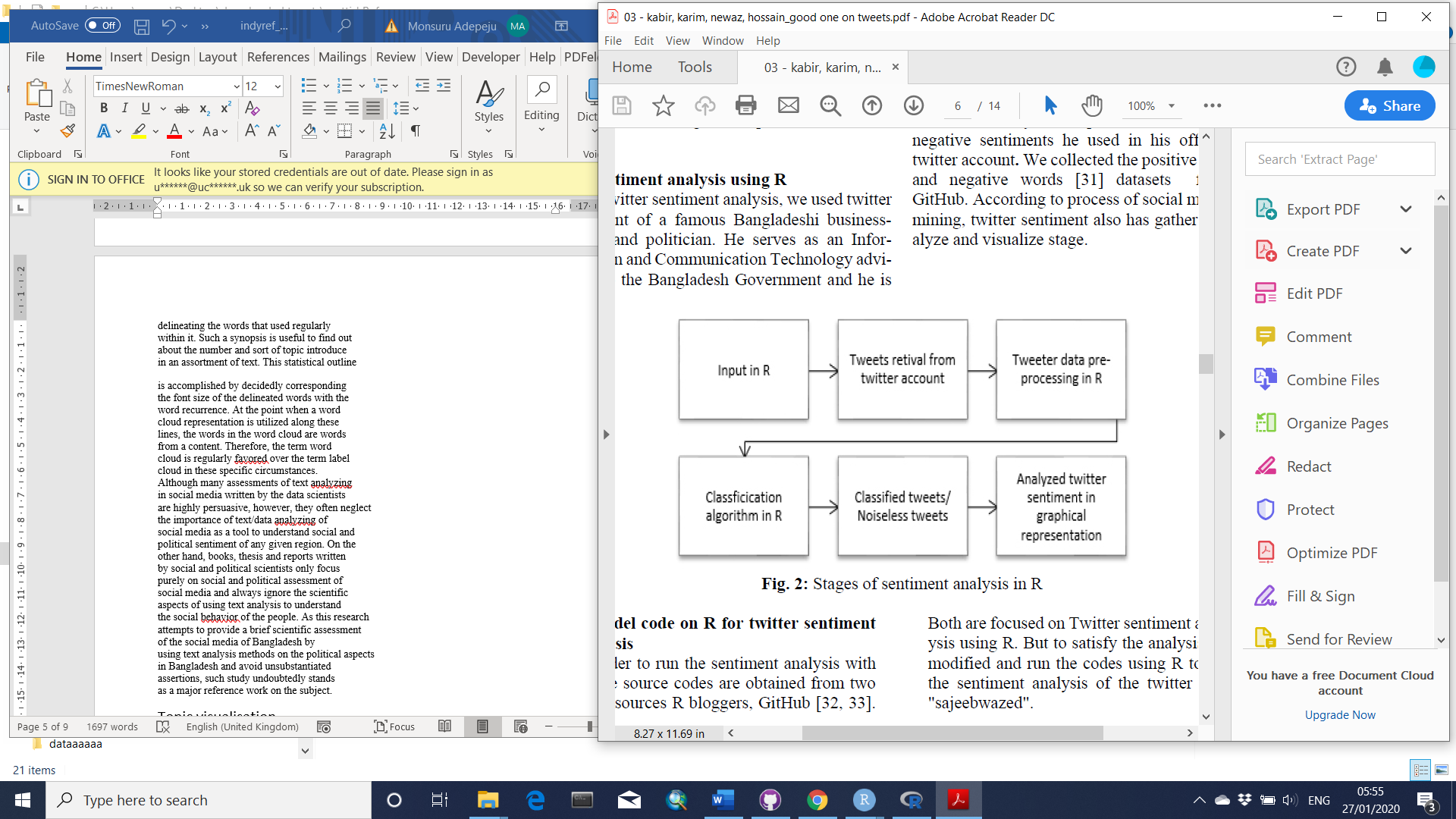
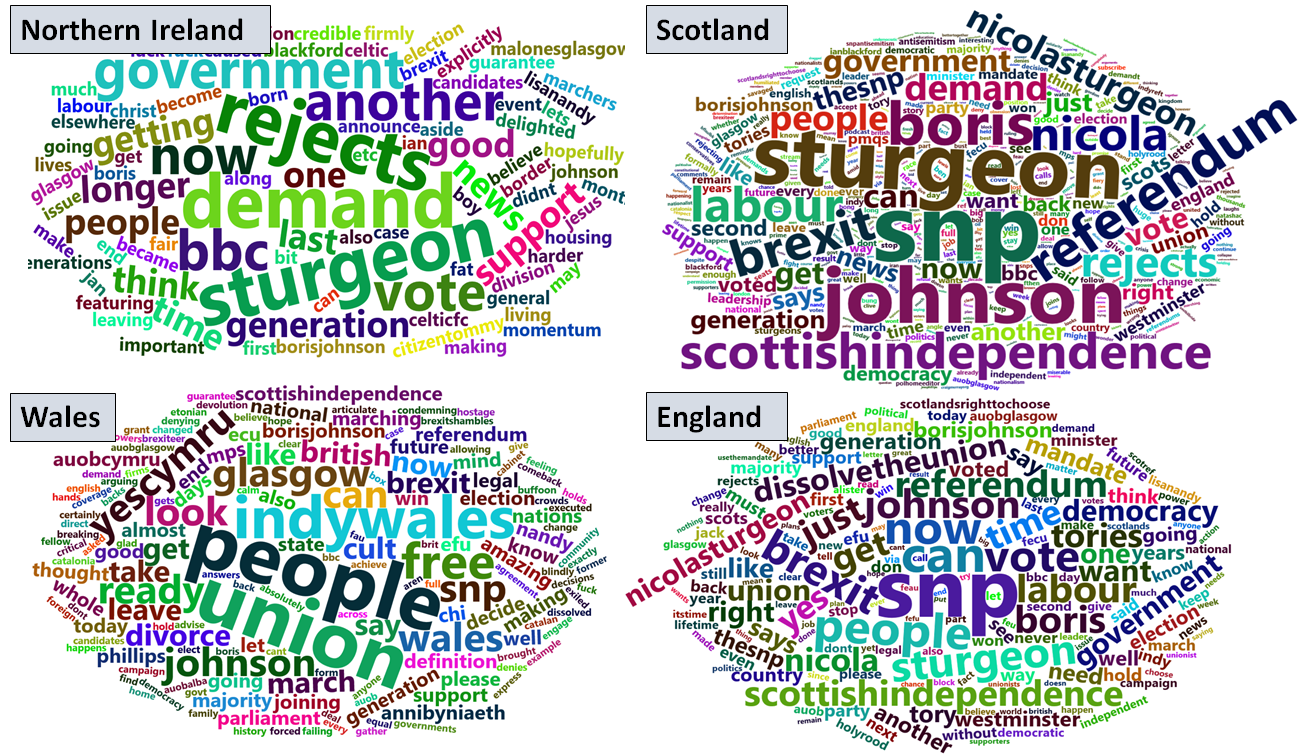
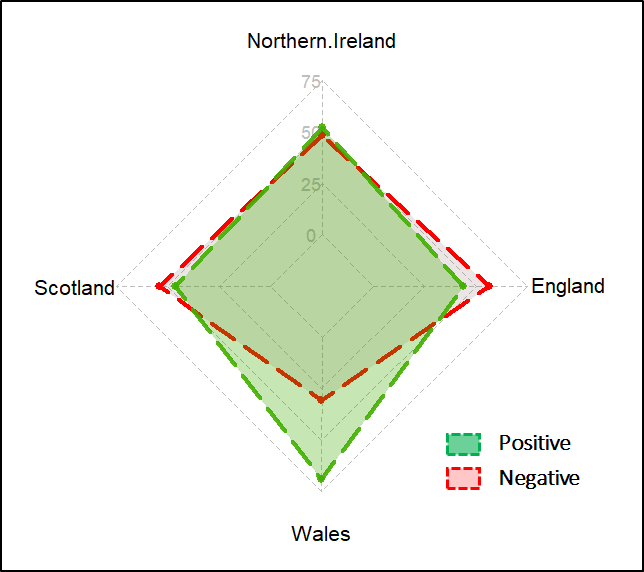
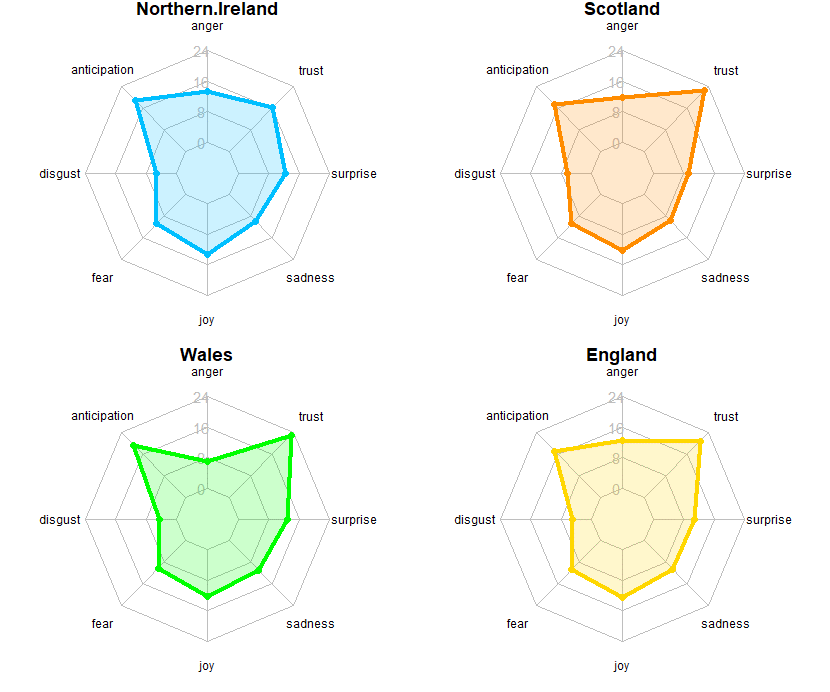


Figure 1 shows that the majority of the tweets (73%) were sent out from the mainland of Scotland, while another 24% from England. Northern Ireland and Wales have small share of 1% and 2%, respectively (Inserted is the map of the UK, showing the relative position of the countries).

Figure 1. Percentage of tweets about Scottish Independence (on Twitter) in the United Kingdom between January 1st and January 31st 2020. Literature review Sentiment analysis is a developing area of Natural Language Processing with research extending from document level characterization [4] to taking in the extremity of words and phrases [5, 6]. Given the character impediments on Tweets, characterizing the sentiment of twitter messages is most like sentence-level sentiment analysis [7, 8]. Be that as it may, the casual and specific dialect utilized as a part of Tweets, and in addition the very idea of the microblogging or social networking domain make twitter sentiment analysis is altogether a different assignment. How well the features and strategies utilized on more structured information will transfer to the microblogging or social media area. Just in the previous year there have been various papers taking a gander at twitter sentiment [9-12]. Different data scientists and researchers have started to investigate the utilization of grammatical features but the results are still mixed. Features regular to microblogging or social networking are likewise normal, yet there has been little examination concerning the handiness of existing sentiment resources created on non-microblogging information. Specialists have likewise started to examine different methods for naturally gathering training information. A few analysts depend on emojis for characterizing their training information [10]. Barbosa and Feng [13] misuse existing twitter sentiment sites for gathering training information or data. [12] likewise utilize hashtags for making training information, however they restrain their examinations to sentiment/nonsentiment grouping. Being "Born outside the universe of PCs" [2], word clouds ended up plainly well known in the specific community oriented websites, for example, photo sharing website Flickr, advertising firm Technorati or social bookmarking website Delicious, that utilize tagging as an ordering strategy [14]. In the meantime, they have developed as a core system of data representation that is applied in a wide range of contexts. One famous application range for tag clouds or word cloud is text outline [15-17]. Here, word cloud is utilized to give a natural and outwardly engaging diagram of a content by delineating the words that used regularly within it. Such a synopsis is useful to find out about the number and sort of topic introduce in an assortment of text. This statistical outline is accomplished by decidedly corresponding the font size of the delineated words with the word recurrence. At the point when a word cloud representation is utilized along these lines, the words in the word cloud are words from a content. Therefore, the term word cloud is regularly favored over the term label cloud in these specific circumstances. Although many assessments of text analyzing in social media written by the data scientists are highly persuasive, however, they often neglect the importance of text/data analyzing of social media as a tool to understand social and political sentiment of any given region. On the other hand, books, thesis and reports written by social and political scientists only focus purely on social and political assessment of social media and always ignore the scientific aspects of using text analysis to understand the social behavior of the people. As this research attempts to provide a brief scientific assessment of the social media of Bangladesh by using text analysis methods on the political aspects in Bangladesh and avoid unsubstantiated assertions, such study undoubtedly stands as a major reference work on the subject. 

Methodology

Topic visualisation

In Figure 2, the ‘Wordclouds’ is used to show the relative importance of the words used in the posts across each country. The bigger and bolder a word appears, the more often it is mentioned in the posts and the more important it is. Dominant words, such as “Indiref2”, “Scotland”, “Scottish”, “independence”, and all hashtags have been filtered out in order to enable clearer representation. There are similarities and differences between the four countries. Words such as “Brexit”, “Boris”, “Johnson”, “Sturgeon”, etc. can be seen to be of high importance across all countries. These are words that are directly connected with the discussion of a 2nd referendum in a near future. For example, the Prime Minister Boris Johnson has just officially rejected the request from the First Minister Nicola Sturgeon to grant the Scottish people another referendum (<https://www.telegraph.co.uk/politics/2020/01/14/boris-johnson-officially-rejects-second-independence-referendum/>). In terms of the differences, words such as “indywales” and “union” can be identified in Wales, and words such as “snp” and “sturgeon” can be identified in Scotland. These are words of significance within the political context of each country. For example, “Indywales”, is a word that has been attributed to the rising nationalist sentiments in Wales for the past few years, while the word “snp” is an acronym for the Scottish National Party.  Figure 2. Importance of words Mapping the sentiment of tweets I used two sentiment lexicons in order to determine the type of emotion expressed in a tweet. First, is a polarity classifier and the second is the emotion classifier. The polarity classifier categorizes words in a binary fashion into positive and negative sentiments. The outcome of polarity classification is represented in Figure 3, which nicely compares the percentage of tweets with positive and negative sentiment for each country. In Wales and Northern Ireland, the majority of people expressed positive sentiments in their post, with 69% and 51%, respectively. This is in contrast to England and Scotland where the majority expressed negative sentiment, with 52% and 53%, respectively.  Fig. 3 Polarity sentiment (%)  Figure 3. Emotion sentiment (%) Second, the emotion classifier categorises words into several emotional statuses, allowing a broader insights into the emotion behind the posts. The categories of emotions include ‘anger’, ‘anticipation’, ‘disgust’, ‘fear’, ‘joy’, ‘sadness’, ‘surprise’, and ‘trust’. Overall, the patterns of the sentiments are comparable across all the four countries (see Figure 4). ‘Anticipation’ and ‘trust’ are two most expressed sentiments, with Wales showing slightly higher percentages in both categories. ‘Disgust’ is the least expressed sentiments in all the four countries. More analysis is needed in order to provide the justifications for these observed patterns, and their association with the polarity of sentiment in Figure 3.

CONCLUSION Whilst the polarity of sentiment reveal that tweets originating from both England and Scotland have comparable percentages, with majority showing negative sentiments, the same is true between Northern Ireland and Wales, but with majority showing positive sentiment. The variances in the size of tweets with Wales and Northern Ireland having significantly small number of tweet on the subject of Scottish Independent. However, the content of the tweets remain similar in the most part.

Twitter provides a valuable source of diverse range of political sentiments subject on political subjects. To date however, sentiment analysis of the Twitter data have focused on the entire United Kingdom in unison, minimizing the ability to reveal the variation of those political sentiments across the constituent nations, namely the England, Scotland, Wales, and Northern Ireland. The nation-based analysis has never been more necessary, when nationalist sentiments are on the rise across each nation, even in Northern Ireland which voted overwhelmingly to remain in the EU. This article provides a glimpse into the variances in the political sentiments about the Scottish referendum prior to the UK exiting the EU on the 31st January 2020. Most relevant.

5. Conclusion and Future Scope

Only 12.18% of people of the total population in India have communicated using English language and most of the common people express their opinions using native language [30]. Therefore analyzing sentiment using the only English language isn’t worthy of every time. Our research will be more robust when we will implement the lexicon of different native languages used in India (e.g. Hindi, Gujarati etc.) [21]. Furthermore, we will also extend our research using Latent Dirichlet Allocation (LDA) based topic model and also identifies the correlated topic across a different category for the event [26]. 6. Declaration The authors declare and solemnly affirm that this research has neither been funded by any political or religious groups nor are the authors in any way affiliated to any institutions with direct or indirect access to groups with biased interests. This research work has been carried out exclusively and independently by the authors in the interests of technology and progress of science in sentiment analysis and related fields.