IS SOCIAL MEDIA CHANGING THE ENGLISH LANGUAGE?

COMP6234: DATA VISUALISATION COURSEWORK (2017-2018)

Richa Ranjan
University Of Southampton
rr2n17@soton.ac.uk

ABSTRACT

It is always interesting to gain an insight into how the English language is changing every day, and also, how much of these changes are influenced by the use of social media. This report aims at pointing out the trends on some of the social media platforms, and how these contribute to new inclusions in the language.

INTRODUCTION

English - the global language, is changing every day to accommodate new developments. Undoubtedly, social media is one of the driving forces towards this evolution of language. And because the language we use to communicate with each other tends to be more malleable than formal writing, the combination of informal, personal communication and the mass audience afforded by social media is a recipe for rapid change.[1] Be it introducing fashionable acronyms, or new meanings of words, new words or even new phrases altogether, social media has pretty much "*been there done that*"!

METHODOLOGY

Various social media platforms like *Facebook*, *Twitter*, *Reddit*, *Pinterest* and many others have contributed to the trends that lead to language changes. Here, we will be talking about some of the catchy words/phrases or acronyms, popularized by social media platforms to an extent where they end up getting etched in the books, for future generations to use. One of the most authentic references for people to look up for any English word/phrase is the **Oxford English Dictionary** or what is popularly known as the **OED**. Every quarter, the OED publishes latest additions made to the dictionary. In this report, we would look into the following data to find out how many of these additions are influenced by their usage on social media:

- Most popular Buzzwords on *Twitter*
- *Google trends* data
- Frequency of occurrence of these buzzwords on *Reddit*

The most popular words across these platforms are searched for in the Oxford Dictionary. This would serve as an evidence that a lot of the OED's new additions are actually inspired by the social media trends.

0.1 DATA COLLECTION

As this is a cross-platform analysis, data from multiple sources were collected:

- **Twitter:** The tweets of three different years were collected, namely 2009, 2016 and 2017. The timelines being apart were an advantage as this helped in a clear examination of the word popularity trend. The tweets

from 2009 ranged for up to 90 days, September to December. 2016 tweets ranged only for a week, and the 2017 data ranged for a month, namely October, 2017.

- **Google Trends:** Another source would be the Google Trends data. It is anonymous, categorized (determining the topic for a search query) and aggregated (grouped together). This allows us to measure interest in a particular topic across search, from around the globe. Google Trends provides statistics on search volume, as a proportion of total search for a given term. Again, the most popular words collected from Twitter were searched here. The data was generated in the form of csv, and it showed the relative occurrence of those words compared to others.

words added to the dictionary during that time.

0.2 DATA CLEANSING AND PROCESSING

As most of the data was collected from social media platforms, they were un-formatted long texts. Data collected from Twitter looked like what is shown in figure 1. Evidently, this data would not be proper to carry out the analysis I intended to. The basic sanitization like eliminating hash tags, or similar special characters were mandatory. Additionally, the tweets had to be parsed because specific words that I was looking for, were parts of longer sentences. In order to parse the tweets and take only the most popular words, I used some python codes, a snippet of which is shown below:

```
Ok today I have to find something to wear for fri cuz I don't think I have time any other day this week.. I'm thinking all bl
I am glad I'm having this show but I can't wait till it is over so I can rest and stop worrying !!      2010-03-15 16:53:44
Honestly I don't even know what's going on anymore      2010-03-15 16:52:59
@LovelyJ_Janelle hey sorry I'm sitting infront of this sewing mching ... @Iam_MarkyMark should be calling u soon :)     2010-
Sitting infront of this sewing machine ... I don't feel like doing this ... I'm tired and feeling lazy. And bored .. And lone
Finally home alone!!!! Time to sing and dance around the house  2010-03-15 11:49:07
Sadly I don't own a pic of me where I'm not at the bar, I don't have a drink in my hand or I'm not slightly intoxicated.. Lol
```

Figure 1: Twitter 2009 data

- **Reddit:** Similar to the Google-ngram feature, Reddit hosts an ngram service which is called the "**Project fivethirtyeight**". To get a sense of the language used on Reddit, every comment since late 2007 has been parsed and this tool is built, which enables users to search for a word or phrase to see how its popularity has changed over time. The top buzzwords were entered to check their frequencies of occurrence over the specified time period. The data was then generated in the form of a csv file for further analysis.
- **OED data:** The Oxford English Dictionary's quarterly updates were collected over the last few years. This list provided all the

```
if sampleword in lineitems[0]:
   if re.search('[a-z]', lineitems[0]) is not None:
       if dict1[sampleword] is None
       or dict1[sampleword]==0:
        dict1[sampleword]=int(lineitems[1])
       else:
        dict1[sampleword]=dict1[sampleword]
                         +int(lineitems[1])
          break
       else:
           break
```

The words could now be stored in an excel sheet, that would further help in building visualisations specific to these words.

Apart from Twitter data, the trends from Google and Reddit also showed the occurrences in scientific notations, or in percentages. All of these had to be stored in a uniform notation across all records.

Also, the data from the Oxford dictionary website was in the form of lists, published every quarter. All these words were collected and put into one excel sheet for visualizing. From this, the buzzwords collected across all platforms were separated and marked with the year it got added to the OED. This would help to show that if, in a given year, a particular word was popular across the social media platforms and it got added to the dictionary around the same time, it could be inferred that its popularity across internet has been a driving force for its inclusion in the English dictionary.

0.3 DATA STORY

Social media has made communication easier than before. But how has it affected the English language? This is what my story would present.

Back in the early 2000s, words like "*selfie*" or "*unfriend*" had no existence in our communications. Evidently, that's not the case anymore! While new words have been a regular course in the English language, social media has accelerated the process. Let us take an example to substantiate this story. Let's consider one of the most popular words across all social media platforms, "**selfie**". We will navigate through all the platforms to look for this word and try to analyze when it started gaining popularity.

(1) In the **Twitter** data, we see that in 2009, there was no such word called *selfie*. This shows that the word was not much popular then, and was not a part of the Oxford Dictionary too. Going ahead, the *tweets* from 2016 shows that *selfie* was mentioned for about 827 times in a week's time. Also, by 2017, the usage rate increased up to 8625 times in a month! This shows a significant increase in it's frequency.

(2) On **Reddit**, the project "Fivethirtyeight" showed that *selfie* was mentioned about 0.0008% of times in 2013, and then 0.00209% in 2014. This is the period when the word got acceptance into the OED.

(3) On **Google Trends**, the relative interest of the word *selfie* was over 700 times, when compared to that of about 29 other words, in the last decade.

(4) Finally, from the **Oxford** report, it is observed that the word *selfie* was added to the OED in 2014.

It can be inferred from these reports that around 2008-09, the word *selfie* was not in use(or very little use in some cases). The trends show a huge increase in its occurrence around 2012-2013 across all social media platforms. The word's enormous popularity is evident from the fact the it was declared as "**Word of the Year**" by Oxford in 2013! Following that, the OED officially accepted it as a word in 2014. Thus, it would be safe to say that the trends in social media led to the official acceptance of the word, as part of the English Language. Similarly, there are many words that gained acceptance into the OED followed by their immense popularity over the internet.

0.4 DATA VISUALISATION

While there could be several examples to present this data story, I chose the following visualisations, that in my view are the best fits for my analysis:

(1) **Word Cloud:** To analyze Twitter buzzwords, I needed a graph that would highlight the selected words. A word cloud is a visual representation of text data, typically used to depict keyword tags on websites, or to visualize free form text. This was the most appropriate visual as it was a direct representation of the actual words, rather than their usage statistical data. An example of a word cloud visual can be seen in figure 2.

(2) **Line Graph:** To represent the frequency of usage of words on Reddit, the Project *Fivethirtyeight* represents it in the form of line trends. A Line Graph is used to compare changes over the same time period for

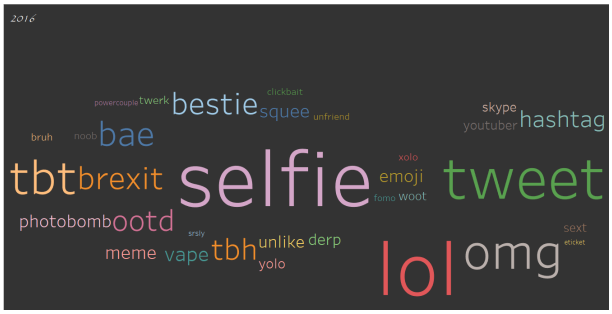

Figure 2: Twitter Buzzwords in 2016

more than one group. As the occurrence frequency is shown over a period of time (a decade in this case) for multiple words, Line Graph is the best choice. Frequency of few buzzwords are shown in figure 3:

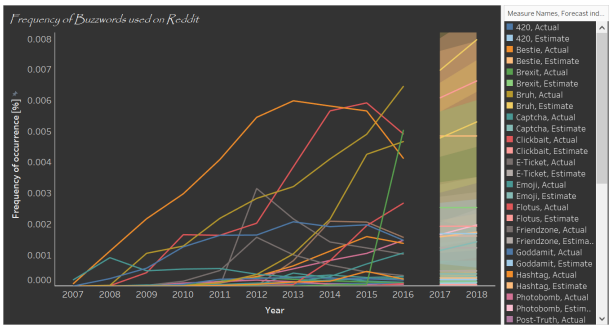

Figure 3: Reddit Buzzwords Frequency

(3) **Bar Graph:** Google Trends show a normalized or relative level of interest over time for a prospective keyword phrase. It also allows us to compare the level of interest among potential target phrases. In this case, I tried to present the Google Trend usage for a set of words. As this comparison is a numerical data, a Bar Graph is the most appropriate visual to see a clear difference in usage frequency, as shown in figure 4:

(4) **Network Graph:** Every year new words are added to the Oxford English Dictionary. In order to show some selected words with their addition dates, the graph shown in figure 5 is a clear representation. A continuous line graph or a bar graph would not have been a good choice as a single

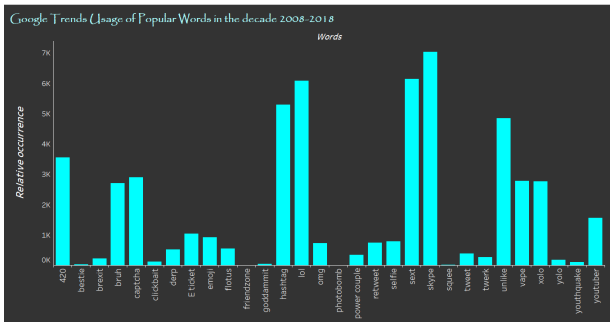

Figure 4: Google Trends data

value (date, in this case) would be unclear to highlight. The circles in the graph shows distinct year for each word.

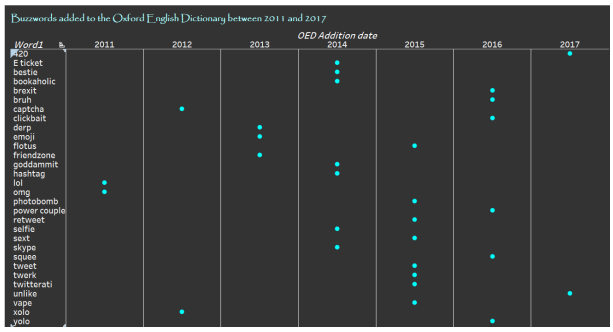

Figure 5: OED word addition dates

RELEVANT THEORIES

The following theories were considered for building the visualisations.

(1) **Gregory's v/s Gibson's Theory of Perception:** The much talked about theory proposed by Gregory is the top-down processing, which states that perception is a hypothesis based on prior knowledge. Contrary to this, Gibson proposed the bottom-up processing, that stated "what you see is what you get".[2] I tried to follow this theory by taking into account factors like *Relative brightness* (a contrasting background in all the visualisations) and *Relative size* (words sized in word cloud according to frequency of occurrence).

(2) **Sanocki and Sulman's experiment:** The experiment stated that colour patterns are better remembered when they are harmonious and also when there are fewer colours. In my visualizations on Google Trends and Oxford data, I have used as few as two colours to present my story. Also, as this experiment stated that colour differences between content and background helps focus the attention on the content, I have used a contrasting dark grey background and the contents are usually contrast like white or aqua.

(3) **Gestalt Theory:** The theory suggests that elements are grouped into 'patterns' to make sense of the world. I have tried to consider the basic elements of this theory i.e. *Emergence* (by presenting a simple bar graph for Google Trends), *Reification* (the entire dataset would be huge, so I have presented roughly 30 words in all my visuals), *Stability* (I presented the word cloud first, and then the Line trend) and *Invariances* (similar words across all word clouds)

(4) **Cleveland and McGill's Theory:** They use the concept of "elementary perceptual tasks" to refer to how we visually-mentally process graph elements using *Position along a common scale*, *Positions along nonaligned scales*, *Length, direction, angle, Area, Volume, curvature, Shading* and *colour saturation* as the elementary tasks. These tasks have been considered while making the above visualisations.

(5) **Tufte's Design Principles:** The theory puts forward concepts like *Graphical excellence, Graphical Integrity, Design Aesthetics* and *Superfluous Information.* In my visualizations, I have restrained from using "the lie factor" and also, "visual clutters" have been taken care of. Also, I have removed irrelevant information to improve identification, by removing words which are not in discussion in this context. As suggested by the theory, I have not used multi-dimensional visuals for one-dimensional variables.

(6) I have also used interactive graphics, as suggested by **Schneiderman**.

CONCLUSION / RESULTS

In conclusion, I would like to state that, from the introduction of new words to new meanings for old words, to changes in the way we communicate, social media is making its presence felt.While there is no fixed numerical evidence about new additions every year, it would be safe to infer that most of the words included in the OED in influenced by social media.

LIMITATIONS AND FUTURE WORK

Given below are few of the limitations of this story:

- There was very little numerical data, and consequently, it was difficult to visualise, because a lot of pre-processing was needed for text data.
- I would be able to show interactivity in a better way if I could present one complete dashboard instead of three different visualisations in the HTML page.
- The size of visualisations were a big concern for me. In order to adjust them in a journalistic article format, i had to resize the visuals in a way that there are no scroll bars required. This led to empty spaces around the visuals.
- The HTML view works well with Google Chrome browser. The text is not properly aligned when opened with Internet Explorer and Firefox.

REFERENCES

[1] https://blog.oxforddictionaries.com/2014/06/18/social-media-changing-language/

[2] https://www.simplypsychology.org/perception-theories.html

[3] https://www.theguardian.com/books/

[4] http://www.bbc.co.uk/news/