Extending Jensen Shannon Divergence to Compare Multiple Corpora

Jinghui Lu[1], Maeve Henchion[2], and Brian Mac Namee[1]

[1] School of Computer Science, University College Dublin, Ireland
[2] Teagasc Food Research Centre, Ireland

Abstract. Investigating public discourse on social media platforms has proven a viable way to reflect the impacts of political issues. In this paper we frame this as a *corpus comparison* problem in which the online discussion of different groups are treated as different corpora to be compared. We propose an extended version of the *Jensen-Shannon divergence* measure to compare multiple corpora and use the *FP-growth* algorithm to mix unigrams and bigrams in this comparison. We also propose a set of visualizations that can illustrate the results of this analysis. To demonstrate these approaches we compare the Twitter discourse surrounding *Brexit* in Ireland and Great Britain across a 14 week time period.

1 Introduction

Social media platforms—such as Twitter, Reddit and Facebook—have dramatically changed the way that people communicate and form their opinions on issues that are important to them [23]. The massive volume of relatively easily accessible digital content generated on these platforms (Twitter alone, for example, has 320 million monthly active users[3]) presents a compelling opportunity to harvest and analyse the opinions of the public on important issues [1].

Many interesting questions that can be answered by analysing data from social media platforms amount to comparing how the opinions of specific groups differ and can be framed as a *corpus comparison*. *Jensen-Shannon divergence* (JSD) [13] is a popular mechanism for performing corpus comparison but is limited to comparing pairs of corpora and considering only bigrams or unigrams. We extend the Jensen-Shannon divergence approach to allow comparison of multiple corpora and enable simultaneous analysis of both unigrams and bigrams through the use of the FP-growth algorithm, a popular approach for frequent itemset mining. We demonstrate the effectiveness of this approach through an analysis of the differences in Twitter data relating to *Brexit* arising from Ireland (including Northern Ireland) and Great Britain (excluding Northern Ireland), and across different time periods.

The remainder of the paper proceeds as follows. In Section 2 we survey relevant existing work; Section 3 describes Jensen-Shannon divergence and how we have extended it; Section 4 is a case study of the application of our approach

[3] Data retrieved on July 22, 2017 from `https://about.twitter.com/company`

to analysing the Twitter discussion of Brexit; and, finally, Section 5 summarizes the work and suggests directions for future explorations.

2 Related Work

There are many examples in the literature of researchers harvesting content posted on social media platforms and analysing it to understand public opinion. For example, Conover et al. [5] proposed several approaches to monitoring the political opinions of the general public from Twitter data. Similarly, Bollen et al. [2] analysed sentiments extracted from tweets to reveal how events in the social, political, cultural and economic fields impact on the public mood. Twitter data has also been analysed to reveal the distinctive phrases used by people of different genders [19], and the differences between social protest and counter-protest movements [6]. Twitter has also been used for tracking the levels of disease activity and public concern in the US during the influenza H1N1 pandemic of 2009 [21]. Eiji et al [1] addressed the similar issue of detecting influenza epidemics using Twitter data. Although there are some recognized limitations of the effectiveness of using data from social media platforms such as Twitter for analysing public opinion (for example the narrow demographics of these platforms' users, or the tendency to communicate extreme opinions), this has been shown to be an effective approach to revealing insights [15].

Many of the interesting questions that can be answered by analysing data from social media platforms amount to comparing how the opinions of specific groups differ (for example [19] and [6]). This can be framed as a *corpus comparison* problem [10] in which the posts of the different groups are treated as different text corpora to be compared. Typical approaches to corpus comparison are statistical in nature. For example, the TF-IDF measure [18] can be used to reflect how important a word is to a document in a collection of corpora. It is also possible to apply statistical significance tests across the distribution of words in different corpora. For instance, Leech and Fallon [12] used a χ^2-test to identify whether words are more common in British or American English, and Church and Hanks proposed the Mutual Information (MI) measure [4] which was employed to identify the characteristic vocabulary of corpora [10]. Meanwhile, frequency profiling was later used by Rayson and Garside [17] to extract distinct words over corpora of different domains.

Rather than applying statistical corpus comparison methods simply to tokenized words in a corpus, it can be useful to apply linguistic pre-processing using techniques such as part-of-speech tagging [3], stemming [14], and lemmatization [7]. Weber and Buitelaar [22] adopted a hybrid method that computes a χ^2 value for each term after linguistic processing. Terms of a χ^2 value above a certain threshold value are decided to be relevant to an individual domain. Another widely used hybrid corpora comparison method is Jensen-Shannon divergence (JSD) [13]. For example, Pechenick [16] used JSD to weight the importance of words involved in language evolution. Gallagher et al. [6] used JSD to quantify the divergence between tweets containing the hashtag #BlackLivesMatter and

other tweets including #AllLivesMatter to investigate the differing opinions of protest and counter-protest movements.

Typical approaches to JSD work across pairs of corpora and are based on unigram tokens. We extend these to an approach that can compare multiple corpora and mixtures of unigram and bigram tokens.

3 Extending Jensen-Shannon Divergence

In this section we describe how Jensen-Shannon divergence (JSD) can be used for corpus comparison and how we have extended the standard approach to allow for comparison of multiple corpora and the use of a combination of unigram and bigram tokens.

3.1 Jensen-Shannon Divergence

Broadly *entropy* refers to uncertainty or disorder [9]. Shannon's entropy [20] is a measure of the unpredictability of a state and can be written as:

$$H = -\sum_{i=1}^{n} p_i \log p_i \tag{1}$$

In the text analysis context, Shannon's entropy describes the uncertainty of a text which has n unique words, where the ith word has probability p_i of appearing. In this case, we can use Shannon's entropy as a diversity measure called the *Shannon index*, where higher entropy implies higher diversity (text is more unpredictable) and vice versa.

Kullback and Leibler [11] proposed a statistical measure which estimates the differences between two probability distributions. Given two probability distributions P and Q, the Kullback-Leibler (KL) divergence is defined as:

$$D_{KL}(P\|Q) = \sum_{i=1}^{n} p_i \log_2 \frac{p_i}{q_i} \tag{2}$$

where n is the size of the sample space. In the context of text analysis, n can be regarded as the number of unique words; and p_i and q_i are the probabilities of observing word i in corpora P and Q respectively.

Applying KL divergence directly to two Twitter corpora, however, is likely to raise issues [6]. If a word appears in only one corpus, this divergence will be infinitely large. To avoid this, Gallagher et al suggested implementing the Jensen-Shannon divergence instead, which is a smoothed version of the KL divergence. The JSD was originally proposed by Lin [13] as:

$$D_{JS}(P\|Q) = H(\pi_1 P + \pi_2 Q) - \pi_1 H(P) - \pi_2 H(Q) \tag{3}$$

where $H(x)$ is Shannon's entropy as described in Equation 1 and π_1 and π_2 are weights associated with the two probability distributions P and Q, respectively. Gallagher et al [6] rephrased JSD as:

$$D_{JS}(P\|Q) = \pi_1 D_{KL}(P\|M) + \pi_2 D_{KL}(Q\|M) \tag{4}$$

This solves the issue of infinite divergence by introducing the mixed distribution $M = \pi_1 P + \pi_2 Q$, where π_1 and π_2 are weights proportional to the sizes of P and Q, with $\pi_1 + \pi 2 = 1$. JSD has a useful property that it is bounded between 0 and 1. When comparing two texts, if a JSD score equals 0 this indicates that the word probability distributions in both texts are equal. A JSD score of 1 indicates that there is no word that appears in both distributions [6].

Another advantage is that we can measure the contribution to the divergence of individual words by the linearity of JSD. The contribution of word i to JSD can be calculated by:

$$D_{JS,i}\left(P \,||\, Q\right) = -m_i \log_2 m_i + \pi_1 p_i \log_2 p_i + \pi_2 q_i \log_2 q_i \tag{5}$$

where m_i is the probability of seeing word i in M. Through Equation 5, we can easily find the most indicative words of each corpus by sorting the JSD contributions of each possible word. JSD has been previously used to compare two corpora [6, 16]. We extend this idea so that we can not only compare tweets from two countries but also tweets from different time periods.

We can extend Equation 5 so that it can be applied across multiple probability distributions, which would allow more than two corpora to be compared. The extension of Equation 5 that computes word i's contributions to divergence over multiple corpora is given as:

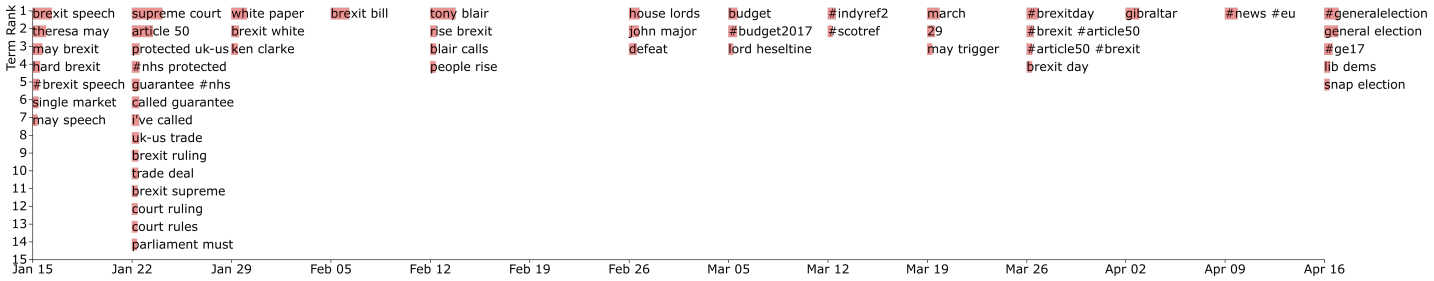
$$D_{JS,i}(P_1 \,||\, P_2 \,||\, ... \,||\, P_n) = -m_i \log_2 m_i + \sum_{j=1}^{n} \pi_j p_{ji} \log_2 p_{ji} \tag{6}$$

where p_{ji} is the probability of observing word i in corpus P_j, and m_i is the possibility of seeing word i in M. Here, M is a mixed distribution of n corpora:

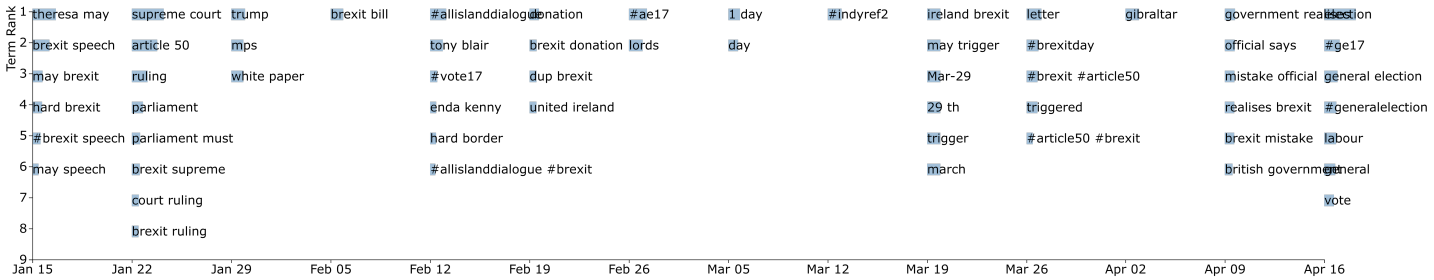
$$M = \sum_{i=1}^{n} \pi_i P_i \tag{7}$$

where, again, π_1, π_2 ... π_n are weights proportional to the sizes of P_1 to P_n, with $\pi_1 + \pi_2 ... + \pi_n = 1$.

By Equation 6, we can compute the contributions of individual words to the JSD divergence value over many corpora. In this study, we apply the extended JSD equation to discover the distinguishing words of tweets from different time periods. We also extend previous approaches to allow unigrams and bigrams to be analysed in parallel and describe the approach to this in the next section.



(a) Great Britain



(b) Ireland

Fig. 1. Scatter plot of top terms from Great Britain (a) and scatter plot of top terms from Ireland (b)

3.2 The FP-growth Algorithm

JSD can be easily applied at both the unigram and bigram levels by considering unigram or bigram tokens in separate applications of the calculations described in the previous section and combining the results. This naive approach, however, leads to an unsatisfactory result where the component unigrams of each bigram will also appear in any list of the most divergent terms. To address this issue we apply the *FP-growth* algorithm[4] [8] to discover all frequent sets of tokens which satisfy a minimal support level (in our implementation a frequency equal to at least the square root of the number of words in the corpus). If a unigram and a bigram are included in the same frequent set of tokens, the unigram can be recognized as redundant and removed. By using FP-growth to filter out redundant information, we can analyse bigrams together with unigrams to give better analysis.

4 Case Study

In this section we present a case study of using our extended JSD approach to compare the Twitter discussion relating to Brexit in Ireland (including Northern Ireland) and Great Britain (excluding Northern Ireland) across different time periods. We are concerned with two questions: (1) how did people's concerns over Brexit change over the time period, and (2) what are the different concerns in relation to Brexit in the Great Britain and Ireland? We first describe how we collected a dataset, then describe a set of visualizations used to present the analysis, and finally the insights which are extracted from the analysis.

4.1 Data Collection

Our dataset was obtained from Twitter using the Twitter Get Search API[5]. We collected tweets relating to Brexit from Ireland and Great Britain over the time period from 15/01/2017 to 23/04/2017. To collect tweets relating to Brexit we specify that a tweet must contain at least one of the search terms "*brexit*, "*post-brexit*", "*hard-brexit*", "*soft-brexit*", "*postbrexit*", "*softbrexit*", or "*hardbrexit*". To separate tweets from Ireland and Great Britain we specify spatial regions through a centre and radius (as allowed through the API). The details are:

- Ireland: latitude: 53.413940, longitude: -7.940989, radius: 300
- GB (south): latitude: 52.674554, longitude: -1.761640, radius: 220
- GB (north): latitude: 56.268001, longitude: -5.185579, radius: 300

Great Britain is divided into two regions, GB (north) and GB (south), with tweets from both combined into a single corpus.

After collecting tweets using these criteria we drop all duplicate tweets and retweets. Our final dataset contained 1,129,754 tweets from Great Britain and

[4] Using the pyfpgrowth package in Python

[5] `https://dev.twitter.com/rest/reference/get/search/tweets`

72,148 tweets from Ireland. Before beginning analysis of this dataset we removed all punctuation (except for # and @ symbols), converted all text to lowercase, and removed stop words. Following this we tokenised separately to unigrams and bigrams.

4.2 The Evolution of Attention

Figure 1(a) shows an illustration of how the top concerns over Brexit of British people changed over the period from 15/01/2017 to 23/04/2017, according to an analysis of collected tweets. We have divided this time period into 14 periods of 7 days, each of which defines a corpus. We apply Equation 6 to compute how much individual unigrams and bigrams contribute to the divergence across these 14 time periods. Then we rank the unigrams and bigrams according to their contribution scores.

As we look at Figure 1(a), each term is represented by a horizontal bar with a width indicating the JSD contribution score. Each term only appears once in the graph in the time period that has the highest possibility of seeing that term. The vertical position of each term represents the rank of the JSD contribution score for the term in a certain period. For example, the term "supreme court" is located at rank 1 for the week starting on January 22. The JSD contribution score for this term is high denoting that it is the most distinct phrase between Jan 22 and Jan 29. To produce Figure 1(a) we select the top 50 bigrams and top 50 unigrams and then use the FP-growth algorithm to remove unigrams that carry duplicate information.

From Figure 1(a), we can see the change of British people's concerns during different time periods. In general, before February, British people were concerned with topics surrounding British Prime Minister Theresa May's speech, Article 50, and the Supreme Court. In contrast, during February and March, the British people's concerns seemed to become distracted by many other issues when considering Brexit, like Budget 2017, and the Scottish independence referendum as evidenced by the presence of terms "#budget2017", "#scotref", and "#indyref2". However, at the end of March, the topics around Article 50 came back to the public sight. Theresa May signed the letter to trigger Article 50 and instigate Brexit on March 29th which also explains the high ranks of phrases "#brexitday", "may trigger", and "#article50 #brexit" at that time. Finally, at mid April, people's focus appears to be dominated by topics relating to the 2017 British general election.

Figure 1(b) shows the most distinctive phrases from Irish tweets over the same time periods. The result shows an extremely similar situation to the British one. Overall, from January 15 to April 23, the focus of Twitter attention to Brexit in Ireland surrounds Theresa May's speech, the triggering of Article 50, the Scottish independence referendum, and the British general election. There are some differences, however, evidenced by the appearance of terms like "united Ireland" and "hard border". In the next section we focus explicitly on analysing these differences.

4.3 Comparing Brexit in Ireland and Great Britain

To compare the differences between the Twitter discussions of Brexit in Ireland and Great Britain across the full time period covered by our dataset we apply Equation 5 to determine the most divergent unigrams or bigrams. We present the results in Figure 2. We list the top 20 unigrams and top 20 bigrams from each of Ireland and Great Britain (again, we remove the unigrams which are included in high ranking bigrams according to the result of the FP-growth algorithm). The length of the bars corresponds to JSD contribution scores with higher values indicating more distinguishing words. A bar to the left (shaded red) indicates that a term is more common in British tweets, while a bar to the right (shaded blue) indicates that a term is more common in Irish tweets.

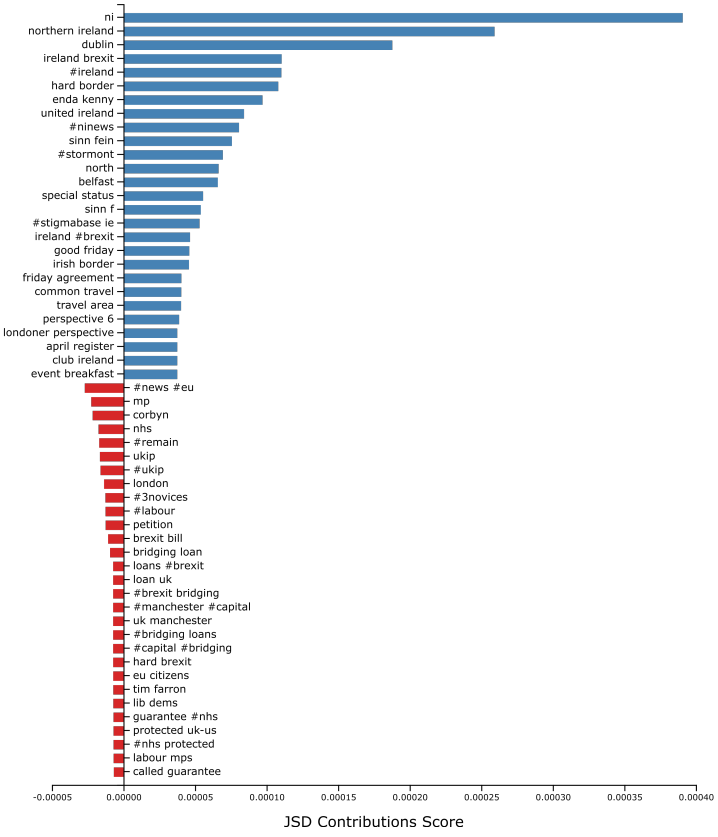
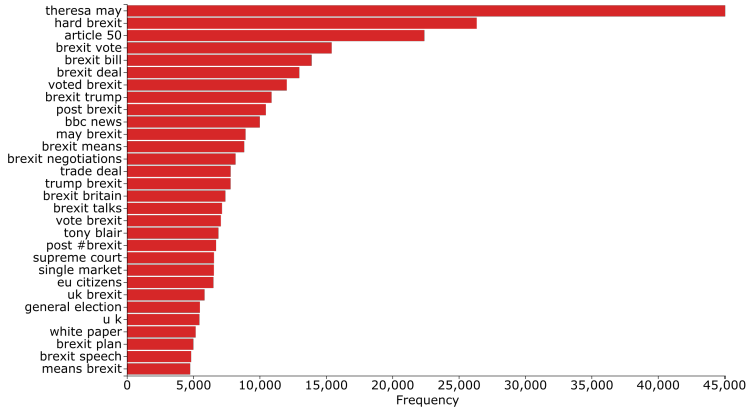
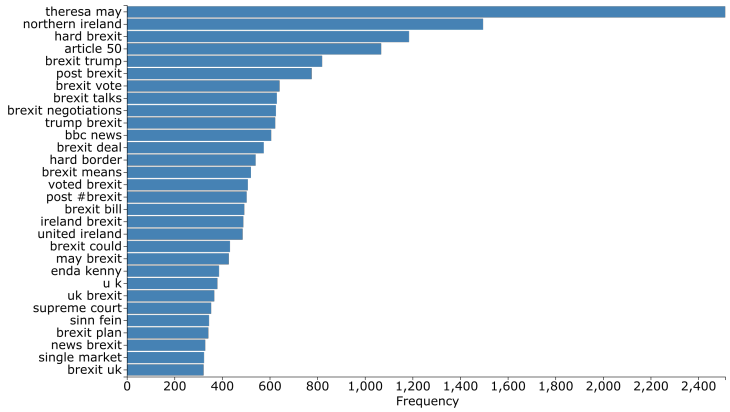


Fig. 2. The most divergent unigrams and bigrams (according to JSD contribution) in British (red and to the left) and Irish (blue and to the right) tweets during the period from 15/01/2017 to 23/04/2017.

From Figure 2 we can see that the top two terms spoken about in Irish tweets, but not in British tweets, are “ni” (an abbreviation for Northen Ireland) and “northern ireland”. This illustrates that the key difference between the concerns regarding Brexit expressed on Twitter by people from Ireland and those expressed by people in Great Britain is a focus on the impact on Northern Ireland and in particular, its border with the Republic of Ireland. We see this echoed in terms like “stormont” (the seat of parliament in Northern Ireland), “hard border”, “sinn fein” (an Irish republican political party), “united ireland”, “good friday”, “friday agreement” (the Good Friday Agreement was a key instrument in peace talks between the Republic of Ireland and Northern Ireland), “common travel”, and “enda kenny” (the Irish prime minister at the time that the tweets were collected).



(a) Great Britain



(b) Ireland

Fig. 3. (a) 30 terms with highest frequency from British tweets and (b) 30 terms with highest frequency from Irish tweets

Conversely, the British tweets seem focused on local issues such as "corbyn" (the British Labour Party leader Jeremy Corbyn), "#ukip" (the Eurosceptic United Kingdom Independence Party), and the "nhs" (a shortened form of the "National Health Service"); and potential impacts of Brexit such as "eu citizens", "hard brexit", and "london".

We can see from Figure 2, however, that the JSD contribution scores are different across the British and Irish corpora—Irish tweets tend to give rise to unigrams and bigrams with higher JSD contributions that British tweets. A possible explanation for this is that it arises because the JSD method looks for unigrams or bigrams that consistently appear in one corpus but rarely appear in the other. Figure 3 shows the 30 most frequent terms in each corpus. These graphs show that the main topics surrounding Brexit (e.g. Article 50, Scottish independence referendum, and Theresa May) are discussed in both British and Irish tweets as evidenced by the high frequency of terms "teresa may", "article 50" and so on. But Irish tweets alone have a set of frequently mentioned border-related topics as evidenced by the highest frequency of term "northern ireland" and so on. In contrast, British tweets do not seem to have a set of frequently mentioned topics that do not appear in Irish tweets. The corpus of British tweets is also much larger than the corpus of Irish tweets and this might also contribute to the relatively low JSD contribution scores for unigrams and bigrams from British tweets.

5 Conclusions

In this paper we have proposed an approach to analysing differences in Twitter discourse of different groups around the same topics using corpus comparison techniques. Specifically, we have used an extended version of Jensen-Shannon divergence coupled with the application of the FP-growth algorithm to merge unigram and bigram analyses. We have also demonstrated how this analysis can be visualized.

We demonstrate this approach through a case study that analyses Twitter discussion of Brexit from Ireland and Great Britain, across different time periods. Through our analysis we can see how concerns over Brexit evolved over the period studied as well as extracting the main differences between the concerns in the two countries—primarily a focus on the impact on the border with Northern Ireland in Irish tweets.

This case study also exposes some of the drawbacks of this approach. For example, it appears that the results of JSD are vulnerable to the effects of spam tweets. The appearance of the phrases "bridging loan", "#brexit bridging", "#Manchester #Capital" etc. reveals that our approach is sensitive to the specific phrases in spam tweets. The reason is simple: if many spam tweets that only appear in one corpus contain very specific terms (e.g. "bridging loan"), then those specific terms will be recognized by the JSD approach as distinguishing. For example, we see in our British corpus (but not in our Irish corpus) various business promotion tweets like "How much can I borrow? - #Manch-

ester #Capital #Bridging Loans #Brexit https://t.co/nDg5ZNKVdf bridging loan, uk, Manchester" that include the distinguishing terms "bridging loan", "#Manchester #Capital" and so on. Though these tweets contain the hashtag "#Brexit" they are actually unrelated to Brexit issues. The JSD approach can be easily hijacked when this happens.

There is also a tension between successfully displaying divergence scores along with frequency in a way that is easy for readers to comprehend. We will address these issues in future work. In future work we will also address how similar techniques can be used to compare corpora that arise from quite different sources—for example online news sources and Twitter.

Acknowledgement. This research was kindly supported by a Teagasc Walshe Fellowship award (2016053).

References

1. Aramaki, E., Maskawa, S., Morita, M.: Twitter catches the flu: detecting influenza epidemics using twitter. In: Proceedings of the conference on empirical methods in natural language processing (pp. 1568-1576) (2011)
2. Bollen, J., Mao, H., Pepe, A.: Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In: ICWSM, 11, pp.450-453 (2011)
3. Brill, E.: A simple rule-based part of speech tagger. In: Proceedings of the workshop on Speech and Natural Language (pp. 112-116). Association for Computational Linguistics (1992)
4. Church, K., Hanks, P.: Word association norms, mutual information, and lexicography. In: Computational linguistics, 16(1), pp.22-29 (1990)
5. Conover, M., Gonalves, B., Ratkiewicz, J., Flammini, A., Menczer, F.: Predicting the political alignment of twitter users. In: Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on (pp. 192-199). IEEE (2011)
6. Gallagher, R., Reagan, A., Danforth, C., Dodds, P.: Divergent discourse between protests and counter-protests:# blacklivesmatter and# alllivesmatter. In: arXiv preprint arXiv:1606.06820 (2011)
7. Green, N., Breimyer, P., Kumar, V., Samatova, N.: Webbanc: Building semantically-rich annotated corpora from web user annotations of minority languages (2009)
8. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. In: ACM sigmod record (Vol. 29, No. 2, pp. 1-12) (2000)
9. Kelleher, J.D., Mac Namee, B., D'Arcy, A.: Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT Press (2015)
10. Kilgarriff, A.: Comparing corpora. In: International journal of corpus linguistics, 6(1), pp.97-133 (2001)
11. Kullback, S., Leibler, R.: On information and sufficiency. In: The annals of mathematical statistics, 22(1), pp.79-86 (1951)
12. Leech, G., Fallon, R.: Computer corpora: what do they tell us about culture. In: ICAME journal, 16 (1992)

13. Lin, J.: Divergence measures based on the shannon entropy. In: IEEE Transactions on Information theory, 37(1), pp.145-151 (1991)
14. Lovins, J.: Development of a stemming algorithm. In: Mech. Translat. & Comp. Linguistics, 11(1-2), pp.22-31 (1968)
15. O'Callaghan, D., Prucha, N., Greene, D., Conway, M., Carthy, J., Cunningham, P.: Online social media in the syria conflict: Encompassing the extremes and the in-betweens. In: Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on (pp. 409-416). IEEE. (2014)
16. Pechenick, E., Danforth, C., Dodds, P.: Is language evolution grinding to a halt? the scaling of lexical turbulence in english fiction suggests it is not. In: Journal of Computational Science, 21, pp.24-37 (2017)
17. Rayson, P., Garside, R.: Comparing corpora using frequency profiling. In: Proceedings of the workshop on Comparing Corpora (pp. 1-6). Association for Computational Linguistics (2000)
18. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. In: Information processing & management, 24(5), pp.513-523 (1988)
19. Schwartz, H., Eichstaedt, J., Kern, M., Dziurzynski, L., Ramones, S., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M., Ungar, L.: Personality, gender, and age in the language of social media: The open-vocabulary approach. In: PloS one, 8(9), p.e73791 (2013)
20. Shannon, C.: A mathematical theory of communication. In: ACM SIGMOBILE Mobile Computing and Communications Review, 5(1), pp.3-55 (2011)
21. Signorini, A., Segre, A., Polgreen, P.: The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. In: PloS one, 6(5), p.e19467 (2011)
22. Weber, N., Buitelaar, P.: Web-based ontology learning with isolde. In: Proc. of the Workshop on Web Content Mining with Human Language at the International Semantic Web Conference, Athens GA, USA (Vol. 11) (2006)
23. Zappavigna, M.: Discourse of twitter and social media: How we use language to create affiliation on the web (2015)