Stylistic features in clustering news reporting: News articles on BREXIT

Abdul Sittar
abdul.sittar@ijs.si
Jožef Stefan Institute and Jožef Stefan Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia

Jason Webber
jason.webber@bl.uk
British Library
London, United Kingdom

Dunja Mladenić
dunja.mladenic@ijs.si
Jožef Stefan Institute and Jožef Stefan Postgraduate School
Jamova cesta 39
Ljubljana, Slovenia

ABSTRACT

We present a comparison of typical bag-of-words features with stylistic features. We group the news articles published from three different regions of the UK namely London, Wales, and Scotland. Hierarchical clustering is performed using typical bag-of-words and stylistic features. We present the performance of 25 stylistic features and compare them with the bag-of-words. Our results show that bag-of-words are better to be used while clustering news reporting at the regional level whereas stylistic features are better to be used while clustering news reporting at the level of news publishers/newspapers.

KEYWORDS

news reporting, topic modeling, stylistic features, clustering

1 INTRODUCTION

The role of content is an essential research topic in news spreading. Media economics scholars especially showed their interest in a variety of content forms since content analysis plays a vital role in individual consumer decisions and political and economic interactions [6]. The content basically refers to the type of language that is used in the news. It is used to convey meaning and it can impact social and psychological constructs such as social relationships, emotions, and social hierarchy [8]. The everyday act of reading the news is such a big area in which small differences in reporting may shape how events are perceived, and ultimately judged and remembered [5].

News reporting across different regions requires methods to find reporting differences. [7] characterize the relationship between the volume of online opioid news reporting and measures differences across different geographic and socio-economic levels. Scholars across disciplines have explored the institutional, organizational, and individual influences that study the quality and quantity of coverage [3].

Features that could classify news reporting across different regions can be adapted to classify the news. A detailed analysis of textual features is performed by [1] where they derived multiple features for creating clusters of news articles along with their comments. These features include terms in the title, terms in the first sentence, terms in the entire article, etc. Multi-view clustering on multi-model data can provide common semantics to improve learning effectiveness. It exploits different levels of features from the raw features, including low-level features, high-level features, and semantic features [16].

Table 1: List of all the stylistic features that are used for clustering.

No.	Feature	No.	Feature
1.	Percentage of Question Sentences	2.	Average Sentence Length
3.	Percentage of Short Sentences	4.	Average Word Length
5.	Percentage of Long Sentences	6.	Percentage of Semicolons
7.	Percentage of Words with Six and More Letters	8.	Percentage of Punctuation marks
9.	Percentage of Words with Two and Three Letters	10.	Percentage of Pronouns
11.	Percentage of Coordinating Conjunctions	12.	Percentage of Prepositions
13.	Percentage of Comma	14.	Percentage of Adverbs
15.	Percentage of Articles	16.	Percentage of Capitals
17.	Percentage of Words with One Syllable	18.	Percentage of Colons
19.	Percentage of Nouns	20.	Percentage of Determiners
21.	Percentage of Verbs	22.	Percentage of Digits
23.	Percentage of Adjectives	24.	Percentage of Full stop
25.	Percentage of Interjections		

The news coverage registers the occurrence of specific events promptly and reflects the different opinions of stakeholders [4]. We take Brexit as an event to be researched on the topic of news reporting differences across the different regions of the UK. On 23 June 2016, the British electorate voted to leave the EU. This event has already been studied following different aspects such as fundamental characteristics of the voting population, driver of the vote, political and social patterns, and possible failures in communication [2, 9]. In this paper, we explore how different stylistic features help in clustering news articles related to Brexit than bag-of-words (BOW).

Following are the main scientific contributions of this paper:

(1) We present a comparison of clustering (using two different textual features: bag-of-words and stylistic features) for news reporting about Brexit in three different regions (London, Scotland, and Wales) of the UK.
(2) We show in our experiments that the bag-of-words are better to be used while clustering news reporting at the regional level whereas stylistic features are better to be used while clustering news reporting at the level of news publishers/newspapers.

2 RELATED WORK

In this section, we review the related literature about topic modelling, and different types of textual features.

2.1 Topic Modelling

Topic modelling is used to infer topics from the collection of text-document. Some techniques used only frequent words whereas

Information Society 2022, 10 October 2022, Ljubljana, Slovenia

Table 2: Total number of news articles about Brexit published in three different regions (London, Scotland, and Wales).

Regions	Newspapers	News articles	Total
London	bankofengland.co.uk	8	4248
	bbc.com	2209	
	dailymail.co.uk	768	
	Independent.co.uk	191	
	inews.co.uk	52	
	metro.co.uk	1	
	neweconomics.org	1	
	rspb.org.uk	8	
	theguardian.com	1167	
	theneweuropean.co.uk	1	
	thesun.co.uk	235	
	cityam.com	3	
	conservativewomen.uk	1	
	dailypost.co.uk	1	
	ft.com	2	
	mirror.co.uk	9	
	raeng.org.uk	1	
	standard.co.uk	20	
Scotland	news.stv.tv	533	533
Wales	gov.wales	3	280
	nation.wales	122	
	Walesonline.co.uk	156	

some use pooling to generate relevant topics and maintain coherence between topics [14]. Topics are typically represented by a set of keywords. Examples of such algorithms are the Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA) and Probabilistic Latent Semantic Analysis (LSA). Clustering-based topic modelling is another solution.

2.2 Stylistic Features

News reporting differences can be reflected through one's speech, writing, and images etc [10, 12]. A language independent features have been used for different tasks of NLP such as plagiarism detection, author diarization. These features considers the text of documents as a sequence of tokens (i.e. sentences, paragraphs, documents). On the basis of these tokens, various types of statistics could be drawn from any language [13]. Stylistic features represent the writing style of a document and have been used for understanding the author writing styles in the past [10]. We use it to explore the clustering of the news articles based on their reporting differences across different regions. Table 1 shows the list of 25 stylistic features used for the development of our proposed clustering of news articles.

2.3 Bag-of-words

A bag-of-words model is a way of extracting features from text. It is basically a representation of text that describes the occurrence of words within a document. It firstly identifies a vocabulary of known words and then measures the presence of known words. Topic modelling is typically based on the bag-of-words (BOW). The essential idea of the topic model is that a document can be represented by a mixture of latent topics and each topic is a distribution over words [11].

3 DATA COLLECTION

We collected news articles reporting on Brexit in the English language from the UK Web Archive (UKWA). The dataset consists of 5061 news articles after pre-processing. Due to the unavailability of news articles from other regions of the UK, we selected only the regions (London, Scotland, and Wales) which have a sufficient amount of news articles. Table 2 presents the number of news articles published from different regions and by different news publishers.

4 METHODOLOGY

The presented research focuses on clustering news articles. To this end, we experiment clustering with the combination of different features observing their performance. Our methodology consists on four steps and compares the performance of stylistic features and bag-of-words in clustering news articles, as shown in Figure 1.

In the first step, we select Brexit under topic and themes on UK web archive[1]. After crawling the list of news articles, we extracted the meta data of news publishers from Wikipedia-infobox. The meta-data extraction process is explained in our previous work [15]. In this process, we extracted the headquarters of news publishers. Due to the unavailability of news articles from other regions of the UK, we selected only the regions (London, Scotland, and Wales) which have a sufficient amount of news articles. In the second step, we perform parsing of the html web pages and extract the body text.

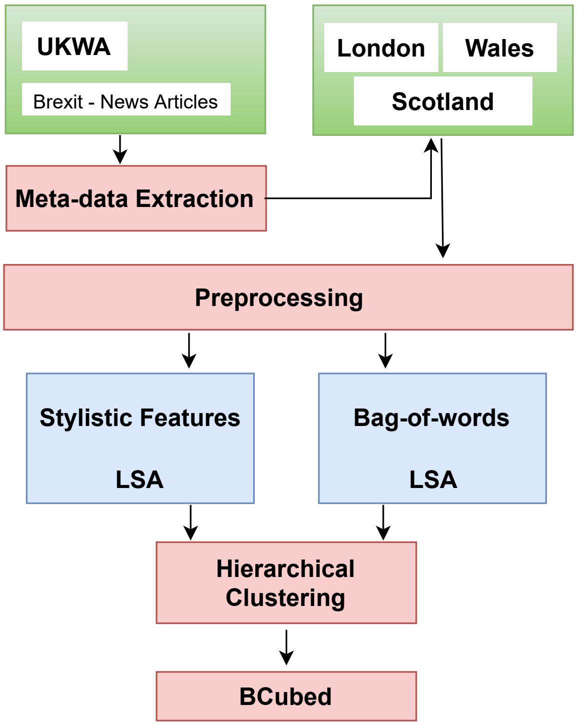

Figure 1: Methodology to clustering regional news using bag-of-words and stylistic features.

[1] https://www.webarchive.org.uk/en/ukwa/collection/910

Since the third step required pre-processing for bag-of-words, we convert the text to lowercase and remove the stop words and punctuation marks. In the third step for the stylistic features, we extract the stylistic features(see Table 1) for all three regions and perform LSA (Latent Semantic Analysis). Similarly, for the bag-of-words, we use the pre-processed text and perform LSA. We also perform LSA on the combination of both types of features. 100 latent dimensions have been used for LSA because it is recommended. We perform LSA and hierarchical clustering using the python library SciPy, and scikit-learn and use the weighted distance between clusters. After performing the LSA, we apply hierarchical clustering and utilize two different types of evaluation measures namely BCubed F1 and Silhouette Scores. For LSA and hierarchical clustering, we use the python library SciPy, and scikit-learn.

5 EXPERIMENTAL EVALUATION

We have performed experimental evaluations using intrinsic (Silhouette) and extrinsic (BCubed-F) evaluation measures. The intrinsic evaluation metrics are used to calculate the goodness of a clustering technique whereas extrinsic evaluation metrics are used to evaluate clustering performance. For extrinsic evaluation, we consider clusters generated by k-means clustering using typical bag-of-words as ground truth clusters. The value of k in k-means clustering ranges from 2 to 20. K-means identifies k centroids, and then allocates every data point to the nearest cluster while keeping the centroids as small as possible. We cannot set the value of k to 1 which means there are no other clusters to allocate the nearest data point.

Silhouette is used to find cohesion. It ranges from -1 to 1. 1 means clusters are well apart from each other and clearly distinguished. 0 means clusters are indifferent, or we can say that the distance between clusters is not significant. -1 means clusters are assigned in the wrong way.

BCubed F-measure defines precision as point precision, namely how many points in the same cluster belong to its class. Similarly, point recall represents how many points from its class appear in its cluster.

- **Silhouette Score:** $S(i) = \frac{b(i)-a(i)}{max(a(i),b(i))}$

where S(i) is the silhouette coefficient of the data point i, a(i) is the average distance between i and all the other data points in the cluster to which i belongs, and b(i) is the average distance from i to all clusters to which i does not belong.

- **BCubed Precision and Recall:**

$$Correctness(i,j) = \begin{cases} 1, if\ L(i) = L(j)\ and\ C(j) = C(j) \\ 0, if otherwise \end{cases}$$

$$BCubed\ Precision = \frac{1}{N}\sum_{i=1}^{N}\sum_{j\epsilon C(i)}\frac{Correctness(i,j)}{|C(i)|}$$

$$BCubed\ Recall = \frac{1}{N}\sum_{i=1}^{N}\sum_{j\epsilon L(i)}\frac{Correctness(i,j)}{|L(i)|}$$

where |C(i)| and |L(i)| denote the sizes of the sets C(i) and L(i), respectively. L(i) and C(i) denote the class and clusters of a point i.

- **BCubed-F Score:** $F = \frac{2 \times BcubedPrecision \times BcubedRecall}{BcubedPrecision + BcubedRecall}$

6 RESULTS AND ANALYSIS

Figure 2 shows the three line graphs. Each graph shows Silhouette scores across a different number of clusters (from 2 to 20) representing different regions of the UK such as Scotland, Wales, and London respectively. Blue and red lines represent bag-of-words (BOW) and stylistic features.

We can see that for all three graphs, the silhouette score of stylistic features is significantly high for all three regions except at one point for Scotland. It means that cohesion is higher and the distance between the clusters is more significant using stylistic features than BOW which is mostly too close to 0. It suggests that these features are better at partitioning news articles into clusters than BOW.

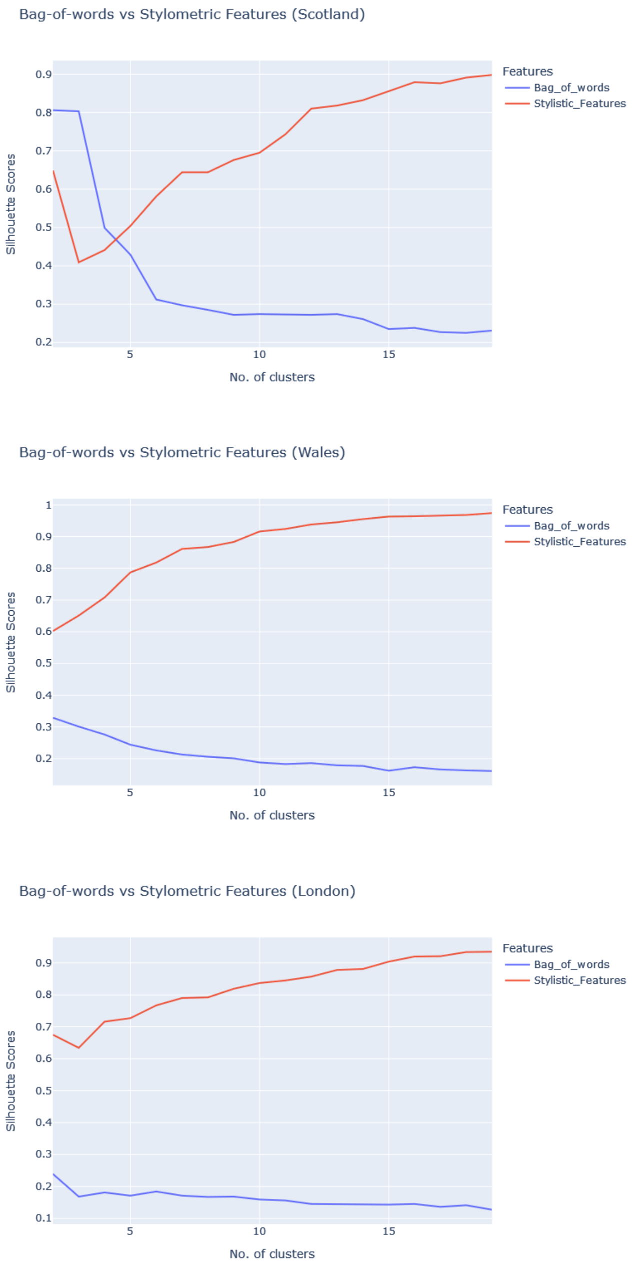

Figure 2: The line graphs represent average silhouette scores across a different number of clusters. The blue line represents the score generated using bag-of-words and the red line represents the score generated using stylistic features. The three-line graphs are generated for three different regions Scotland, Wales, and London respectively.

Table 3: The group of news articles published from three different regions of the UK is considered as ground truth clusters and the Bcubed-F score is calculated using three types of features including bag-of-words, stylistic features, and a combination of both types of features.

No.	Features	Bcubed-F Score
1.	Bag-of-words	0.75
2.	Bag-of-words and stylistic features	0.51
3.	Stylistic features	0.54

Table 4: The group of news articles published from 22 different news publishers of the UK is considered as ground truth clusters and the Bcubed-F score is calculated using three types of features including bag-of-words, stylistic features, and a combination of both types of features.

No.	Features	Bcubed-F Score
1.	Bag-of-words	0.53
2.	Bag-of-words and stylistic features	0.57
3.	Stylistic features	0.66

However, it is insufficient to say that stylistic features are better for news reporting differences at this stage because it is not necessary that the resulting clusters by internal partitioning can be equal to the ones that are based on news reporting differences.

We consider each region (London, Scotland, and Wales) as a ground truth cluster of the news articles published in that region. Table 3 shows Bcubed-F scores when the ground truth clusters were matched with the one that was created using bag-of-words, stylistic features, and a combination of both types of features. Similarly, we consider each newspaper/news publisher shown in Table 2 as a ground truth cluster of the news articles published by that newspaper/news publisher. Table 4 shows Bcubed-F scores when the ground truth clusters were matched with the one that was created using bag-of-words, stylistic features, and a combination of both types of features. The scores using bag-of-words considering regions as ground truth clusters are significantly high (0.75) than stylistic features (0.54) and a combination of all features (0.51). The scores using stylistic features considering newspaper/news publishers as ground truth clusters are significantly high (0.66) than bag-of-words (0.53) and a combination of all features (0.57). The higher scores in regional news reporting suggest that bag-of-words is better to be used for clustering or classification because the newspapers/news publishers report in different styles in a certain region. Similarly, when it comes to classifying or clustering news reporting across different newspapers/news publishers then stylistic features are more useful because the newspapers/news publishers follow a different reporting style.

7 CONCLUSIONS

In this paper, we have presented the comparison of different features observing their performance over clustering news articles. The goal of this work was to investigate the performance of stylistic features and typical bag-of-words. The data consists of news articles about a popular event Brexit that are collected from UKWA. These news articles belong to three different regions of the UK including Scotland, London, and Wales. Our experimental results suggest that bag-of-words are better to be used while clustering news reporting at the regional level whereas stylistic features are better to be used while clustering news reporting at the level of news publishers/newspapers.

ACKNOWLEDGMENTS

The research described in this paper was supported by the Slovenian research agency under the project J2-1736 Causalify and by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 812997.

REFERENCES

[1] Ahmet Aker, Monica Paramita, Emina Kurtic, Adam Funk, Emma Barker, Mark Hepple, and Rob Gaizauskas. 2016. Automatic label generation for news comment clusters. In *Proceedings of the 9th International Natural Language Generation Conference*. Association for Computational Linguistics, 61–69.

[2] Sascha O Becker, Thiemo Fetzer, and Dennis Novy. 2017. Who voted for brexit? a comprehensive district-level analysis. *Economic Policy*, 32, 92, 601–650.

[3] Danielle K Brown and Summer Harlow. 2019. Protests, media coverage, and a hierarchy of social struggle. *The International Journal of Press/Politics*, 24, 4, 508–530.

[4] Honglin Chen, Xia Huang, and Zhiyong Li. 2022. A content analysis of chinese news coverage on covid-19 and tourism. *Current Issues in Tourism*, 25, 2, 198–205.

[5] Elizabeth W Dunn, Moriah Moore, and Brian A Nosek. 2005. The war of the words: how linguistic differences in reporting shape perceptions of terrorism. *Analyses of social issues and public policy*, 5, 1, 67–86.

[6] Frederick G Fico, Stephen Lacy, and Daniel Riffe. 2008. A content analysis guide for media economics scholars. *Journal of Media Economics*, 21, 2, 114–130.

[7] Yulin Hswen, Amanda Zhang, Clark Freifeld, John S Brownstein, et al. 2020. Evaluation of volume of news reporting and opioid-related deaths in the united states: comparative analysis study of geographic and socioeconomic differences. *Journal of Medical Internet Research*, 22, 7, e17693.

[8] Qihao Ji, Arthur A Raney, Sophie H Janicke-Bowles, Katherine R Dale, Mary Beth Oliver, Abigail Reed, Jonmichael Seibert, and Arthur A Raney. 2019. Spreading the good news: analyzing socially shared inspirational news content. *Journalism & Mass Communication Quarterly*, 96, 3, 872–893.

[9] Moya Jones. 2017. Wales and the brexit vote. *Revue Française de Civilisation Britannique. French Journal of British Studies*, 22, XXII-2.

[10] Ifrah Pervaz, Iqra Ameer, Abdul Sittar, and Rao Muhammad Adeel Nawab. 2015. Identification of author personality traits using stylistic features: notebook for pan at clef 2015. In *CLEF (Working Notes)*. Citeseer, 1–7.

[11] Zengchang Qin, Yonghui Cong, and Tao Wan. 2016. Topic modeling of chinese language beyond a bag-of-words. *Computer Speech & Language*, 40, 60–78.

[12] Abdul Sittar and Iqra Ameer. 2018. Multi-lingual author profiling using stylistic features. In *FIRE (Working Notes)*, 240–246.

[13] Abdul Sittar, Hafiz Rizwan Iqbal, and Rao Muhammad Adeel Nawab. 2016. Author diarization using cluster-distance approach. In *CLEF (Working Notes)*. Citeseer, 1000–1007.

[14] Abdul Sittar and Dunja Mladenic. 2021. How are the economic conditions and political alignment of a newspaper reflected in the events they report on? In *Central European Conference on Information and Intelligent Systems*. Faculty of Organization and Informatics Varazdin, 201–208.

[15] Abdul Sittar, Dunja Mladenić, and Marko Grobelnik. 2022. Analysis of information cascading and propagation barriers across distinctive news events. *Journal of Intelligent Information Systems*, 58, 1, 119–152.

[16] Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. 2022. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16051–16060.