Stylistic features in clustering news reporting: News articles on BREXIT

SiKDD 2022

Ljubljana, October 10

Abdul Sittar, Jason Webber, Dunja Mladenic

Jožef Stefan
Institute

Department for
Artificial Intelligence

Problem Definition

- External factors: elections, political alignments, economic issues
- Internal factors: impact social and psychological constructs
- Everyday act of reading the news
- News spreading is directly related to the way the news is reported on any events (individual consumer decisions and political and economic interaction)
- Understanding the framing of news reporting on different topics and in different cultures (political and economic news, military conflict frame, issues and strategy framing by news coverage)

Problem Definition

- Methods are required to identify news reporting differences across regions
- Characterize the relationship between the volume of online news reporting
- News reporting differences about different events are generally inclined towards certain characteristics of newspapers.
- Reflection through its writing and images
- Punctuation marks and the hidden language
- Observe the performance of different features while clustering the news reporting.

Outline

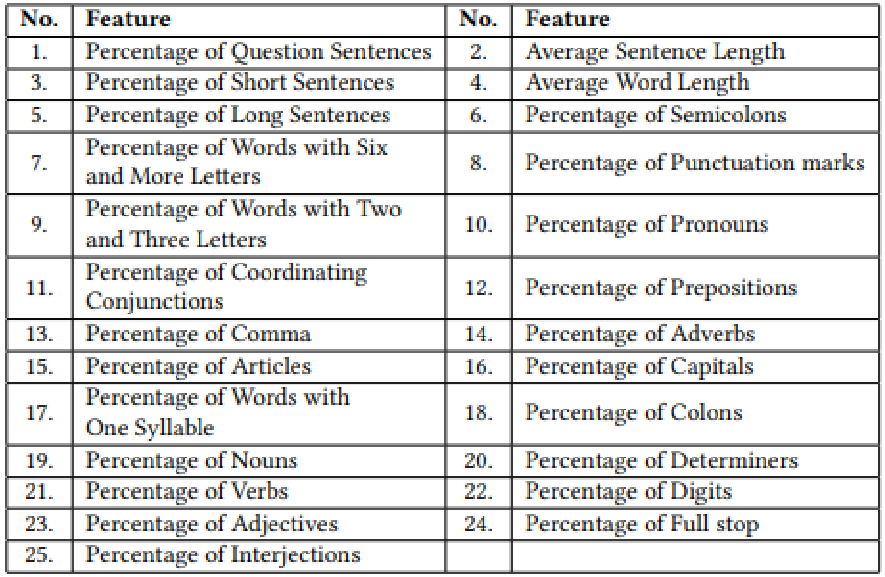
- Introduction
- Contributions
- Data Description
- Methodology
- Experimental Evaluation
- Results and Analysis
- Conclusions

Introduction

- The role of content in news reporting refers to the type of language that is used in the news.
- News agencies/news publishers always want to have more viewership of their content to earn more money.
- The news coverage registers the occurrence of specific events promptly and reflects the different opinions of stakeholders
- 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
- Different aspects such as fundamental characteristics of the voting population, driver of the vote, political and social patterns, and possible failures in communication

Features

- Stylistic features
  - Language independent
  - Plagiarism detection, author diarization
  - Understanding the author’s writing style
  - Clustering of news reporting
- Bag-of-words
  - Simplifying representation in NLP and IR
  - Bag of its words, disregarding grammar and even word order
  - Document classification

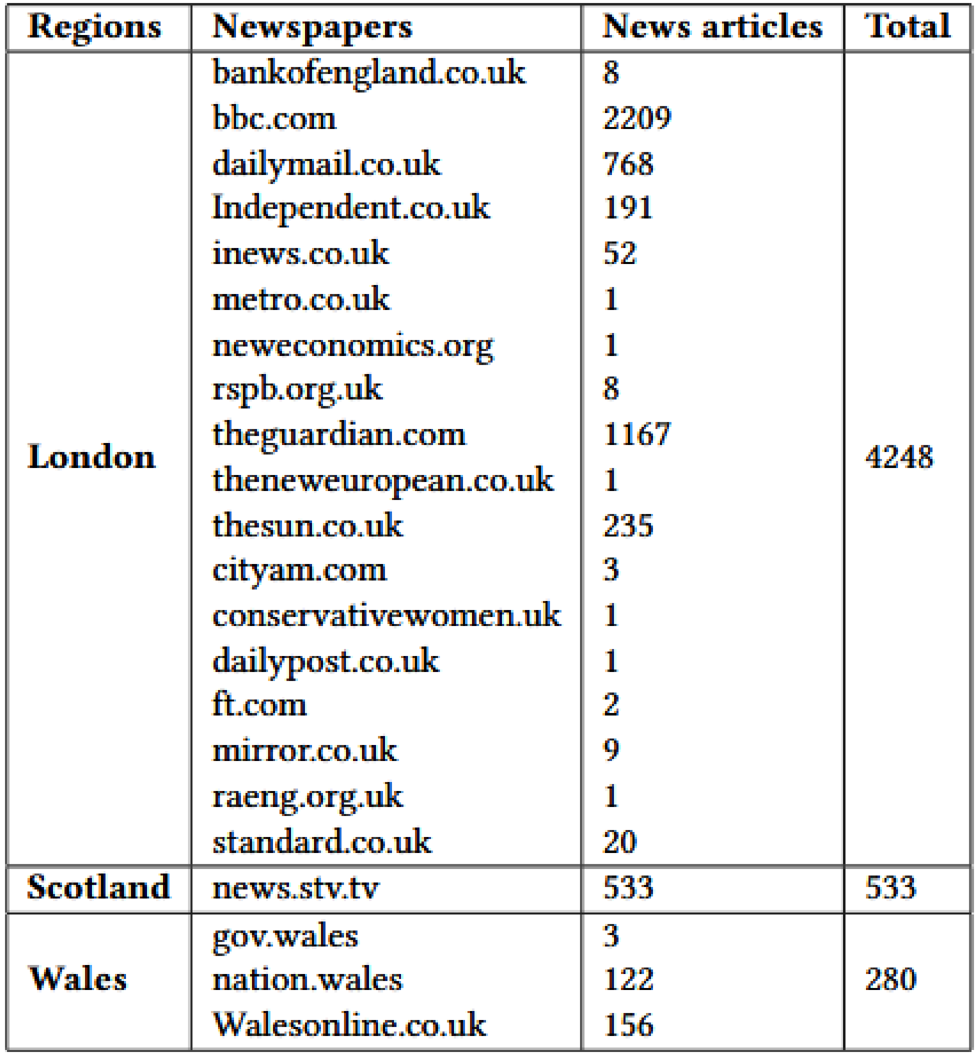


No.	Feature	No.	Feature
1.	Percentage of Question Sentences	2.	Average Sentence Length
3.	Percentage of Short Sentences	4.	Average Word Length
5.	Percentage of Long Sentences	6.	Percentage of Semicolons
7.	Percentage of Words with Six and More Letters	8.	Percentage of Punctuation marks
9.	Percentage 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		

Contributions

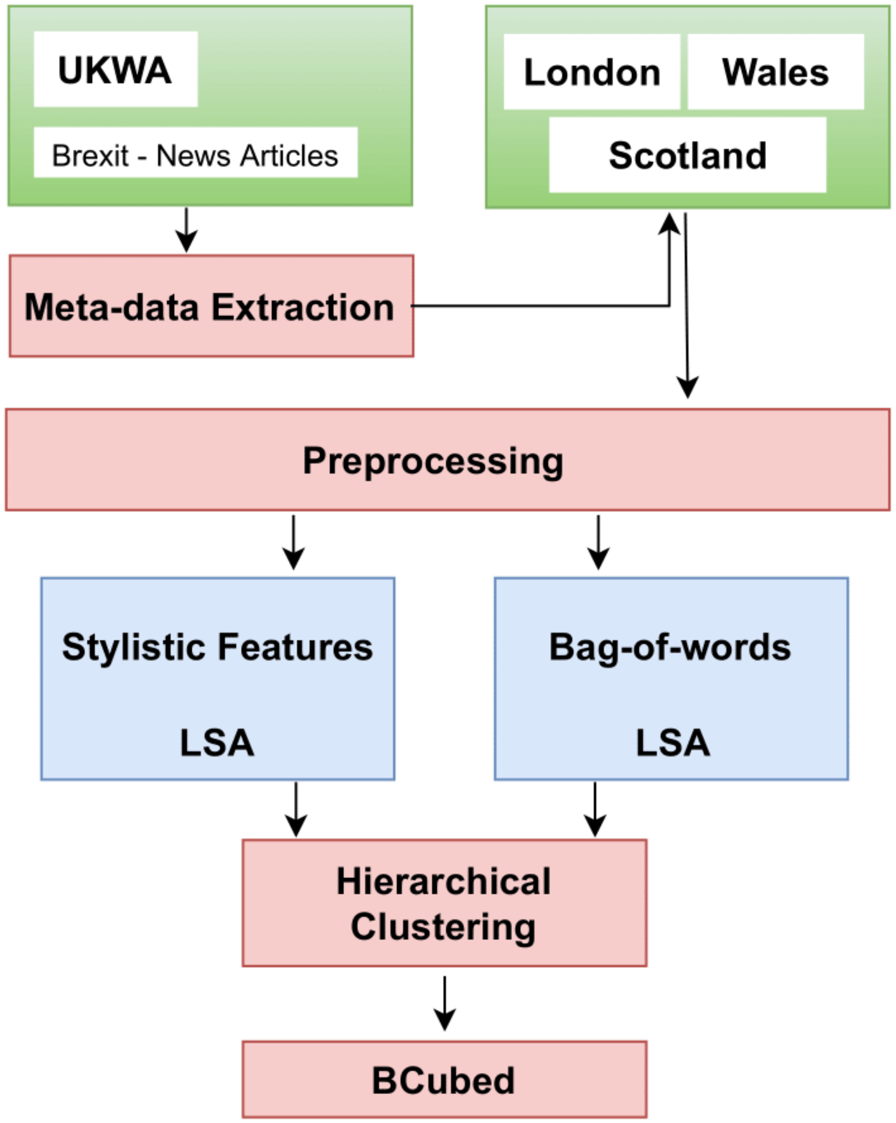
- A methodology to cluster (using two different textual features: bag-of-words and stylistic features) the news reporting.

Data description



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	

Methodology

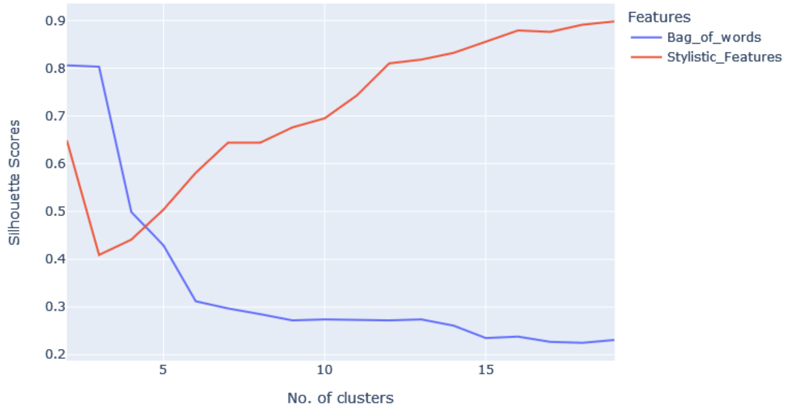


Experimental Evaluation

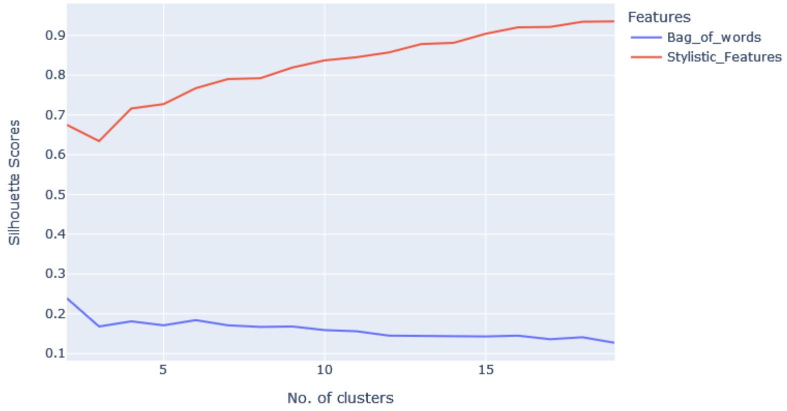
- Silhouette Score
  - Goodness of a clustering technique
  - Find the cohesion
  - -1 to 1
- Bcubed precision and recall
  - Precision as point precision, namely how many points in the same cluster belong to its class.
  - Point recall represents how many points from its class to appear in its cluster.

Results and Analysis

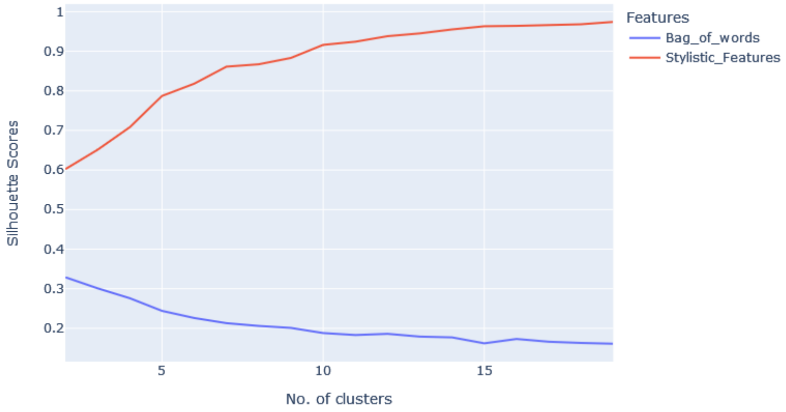
Bag-of-words vs Stylometric Features (Scotland)



Bag-of-words vs Stylometric Features (London)

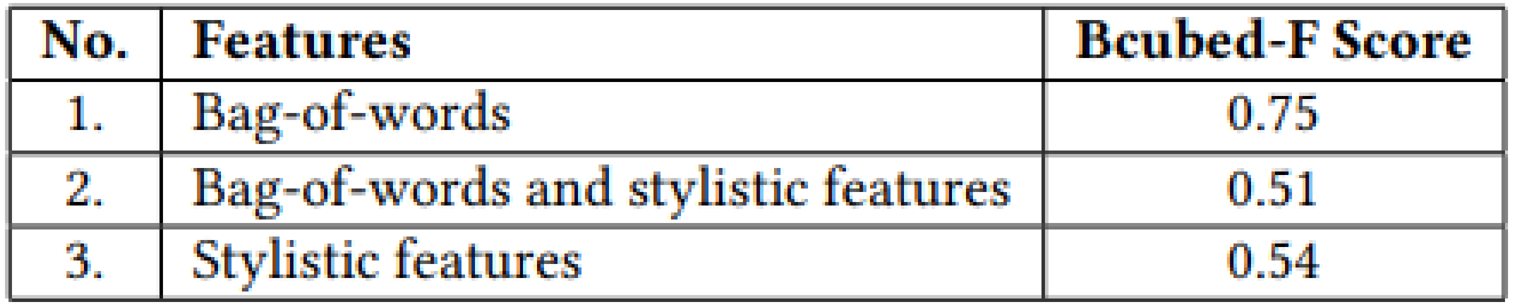


Bag-of-words vs Stylometric Features (Wales)



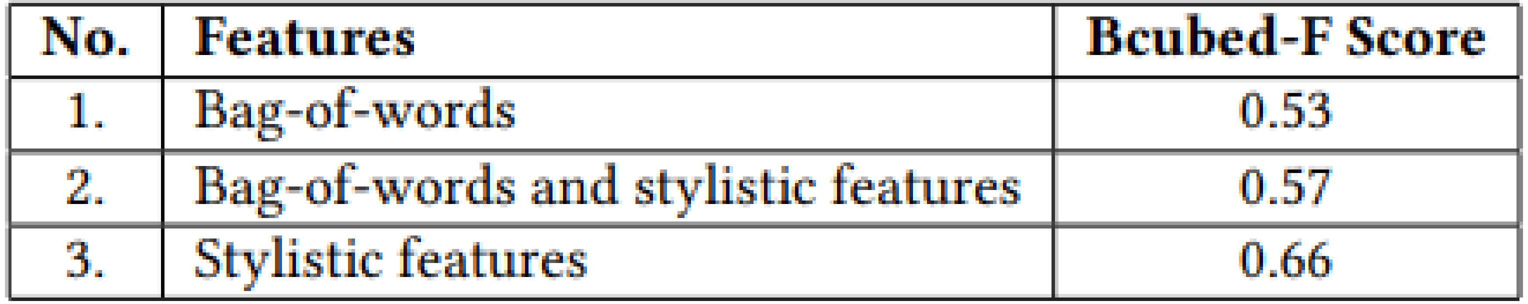
Results and Analysis

- Ground truth clusters across three regions



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

- Ground truth clusters across 22 news publishers



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

Conclusions

- 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 contains news articles about a popular event Brexit that are collected from UKWA
- Bag-of-words are better to be used while clustering news reporting at the regional level
- Stylistic features are better to be used while clustering news reporting at the level of news publishers/newspapers

Future work

- We intend to extend the features to lexical and stylistic features along with already trained models on news data.

Any Questions ?