Polarized Topic Distributions: Comparing Reddit and Irish Times Headlines

Alice Hamberger
alicehamberger@gmail.com
Fiammetta Rosenblatt
fiammetta@online.de
Daria Protsenko
daria.protsenkoo@gmail.com
Nora Svensson Hahr
nora.mitnor@gmail.com

Abstract

In this paper we aimed at comparing the proportion of polarizing topics covered in news headlines in a formal and an informal news context; namely the r/worldnews thread on the social media platform Reddit and the traditional newspaper The Irish Times. We use two distinct topic modeling methods: Latent Dirichlet Allocation (LDA) (Wintrode, 2011) and HDBSCAN clustering (Campello et al., 2015). The results returned by both models demonstrate a larger proportion of polarizing news topics covered within Reddit headlines than in the Irish Times.

1 Introduction

Countering social and political polarization can benefit mental and physical health, as well as help reduce resulting violence and discrimination issues in society (McCoy and Murat, 2019). Understanding how polarization emerges from different news channels can bring insight into developing strategies on how to prevent polarization. Research into this topic has mainly been conducted qualitatively, within the discipline of media studies (Azeem et al., 2019). Although, within the discipline of natural language processing some work has been done in relation to news coverage and headlines (Kroon et al., 2020; Yan & Gao, 2020), to the best of our knowledge no study has specifically focused on the concept of polarization.

Therefore, this research will adopt an empirical approach to add a new dimension in understanding the mechanisms of polarization. It will be asked whether there exists a significant difference in news coverage of polarizing topics in the headlines of the informal Reddit r/worldnews news source as compared to the formal Irish Times newspaper. We will do so using two distinct methods of topic modeling[1]: HDBSCAN clustering and Latent Dirichlet Allocation (LDA). Adopting the political science framing of polarization as a "harsh divide between [...] opposing camps" (Carothers & O'Donohue, 2019), this paper will define polarized topics as given by the existence of (at least) two opposite opinions on a matter.

2 Methodology

2.1 Datasets

The datasets, "Worldnews on Reddit from 2008 to Today" (2016) and "Irish Times - Waxy-Wany News" (2022), were retrieved from Kaggle as csv-files. We analyzed the headlines published in the timespan in which headlines were available from both sources. That is, between January 25, 2008 and November 22, 2016. The Reddit dataset contains around 500,000 headlines and the dataset from the Irish Times contains around 600,000 headlines after fitting it to the given time frame.

2.2 Preprocessing

We only analyzed the headline text, discarding all other information included in the datasets. We tokenized the headlines using regular expressions, removing punctuation symbols, English stopwords, numeric characters, and words shorter than three letters. We subsequently used the WordNet lemmatizer to lemmatize the tokens to retain mainly semantic information. This was the preprocessing used for the latent dirichlet allocation method.

For the HDBSCAN clustering, each headline was subsequently converted to a document embedding. It was decided to manually create document embeddings since the datasets averaged 8.8 words per headline, which was suspected to be too low for off-the-shelf models to produce reliable results. Every headline was represented by a vec-

[1]Code for this project can be found at: https://github.com/norahahr/TMproject

tor calculated through averaging the word embeddings of each word in the headline (Rücklé et al., 2018). Considering the limited size of our dataset, we used the pre-trained Word2Vec model trained on parts of the Google News dataset (about 100 billion words) for the embeddings (Google Code Archive, 2013). The model contains 3 million, 300-dimensional embeddings for various words and phrases.

2.3 Method I: HDBSCAN clustering

The first method consisted of three parts: Optimizing UMAP dimensionality reduction and HDBSCAN clustering, manually labeling the clusters, and manually deciding whether clusters contain polarizing topics. The dimensionality of the embeddings was reduced using UMAP dimensionality reduction (Aggarwal et al., 2001; McInnes et al., 2018). The UMAP algorithm was chosen because of its scalability, ability to preserve global structures in lower dimensions, and because of its common application for data used in density based clustering (Alloui et al., 2020). For clustering, the hierarchical density-based clustering method HDBSCAN was used (Campello et al., 2013; Campello et al., 2015). This algorithm was selected because it does not require the user to specify the number of clusters in the dataset before running the process. Moreover it allows for outliers, and has been shown to perform well on unevenly shaped clusters.

The dimensionality reduction and clustering process were optimized together using a random search algorithm that evaluated 100 hyperparameter combinations. The hyperparameters included the number of components and number of neighbors of the UMAP algorithm as well as the minimum cluster size for the HDBSCAN. The clusters were evaluated on their Silhouette score (Rousseeuw, 1987), the total number of clusters, and the number of embeddings classified as outliers. The optimized clustering model for each dataset was the one with the highest Silhouette score, the lowest number of classified outliers and a number of clusters between 50-200. The cluster range was decided upon after looking at the number of subcategories among the 50 most popular online news sources (Majid, 2022).

Using the optimized clustering, the theme of each cluster was manually labeled based on the 20 most common words in the cluster and 15 randomly sampled headlines. The theme of the cluster was then labeled as either being a polarizing topic or not based on our previously stated definition. Topics such as war and political conflict were for instance labeled as polarizing, but natural disasters or articles on the stock market were not. We also tried to separate clusters that covered polarizing topics from clusters that cover topics that may lead to polarized discussions. For instance, headlines on gun violence in schools may spark a polarized discussion on the right to bear arms. However, most people would agree that gun violence in schools is bad and so there is not much polarization within the topic itself. Therefore, a cluster containing headlines on the topic would be labeled as not covering polarizing material although such headlines could potentially lead to other, polarized, debates.

2.4 Method II: Latent Dirichlet Allocation

We also employed the Latent Dirichlet Allocation (LDA) method to model linguistic themes in the headlines. LDA is a "generative probabilistic model of a corpus" (Blei et al., 2003), and one of the most common tools in topic modeling (Manikonda et al., 2018). For this research, each of the headline corpora is represented as an (overlapping) set of latent topics. In turn, these topics are probability distributions over the words of the vocabulary (Blei et al., 2003).

Using the preprocessed datasets as described above, we create dictionary representations of corpora and additionally filter out common words (occurring in more than 50% of the documents) and rare words (in less than 10 documents). Furthermore, after reviewing the preliminary results of the model, we manually remove some uninformative words like "say" to strengthen our results contextually. Finally, to visualize the topic distribution of each of the news sources, we used the pyLDAvis and Wordcloud libraries.

2.5 Comparing Methods

In this project we decided to use both aforementioned methods in order to compare both thematic analyses and be able to better assess the significance of our results. Both methods differ in their purpose: while Word2vec provides a more local approach (based on word vector similarities), LDA is more global - it does not consider the syntactic structure of headlines and context in which words occur. On the other hand, LDA arguably returns

more easily interpretable results as it is designed for topic discovery (Blei et al., 2003). The output of the model is a direct, fixed number of topics for self-labeling. It is relevant to mention the existence of a hybrid model - lda2vec - that combines the semantic sensitivity of word2vec with the interpretability of LDA topics (Moody, 2016) which was however out of the scope of this paper. Nonetheless, it justifies the comparison of both methods as done in this research.

3 Results

3.1 Method I (HBDSCAN clustering)

A summary of the results found using HDBSCAN clustering can be found in Table 1.

	Reddit	**Irish Times**
Total number of headlines	508319	597058
Total number of labeled headlines	290265	322207
Total number of clusters	154	131
Clusters labeled to contain polarizing clusters	55	23
Sum of headlines in clusters labeled to contain polarizing topics	117893	75657
Relative number of headlines in clusters containing polarizing topics (only counting labeled data)	40.6%	23.5%
Silhouette score of clustering	0.534276	0.558982

Table 1: Results from HDBSCAN clustering

Certain clusters were difficult to associate with any specific topic or theme. We decided to exclude these clusters from the dataset as they did not significantly reduce the number of headlines in the corpora (2.1% of headlines in the Reddit corpus and 5.7% of headlines in the Irish Times corpus).

To judge how representative it was to label an entire cluster as either covering polarizing issues or not, we sampled 15 new headlines from each cluster to see if the headline labels matched the labels of the clusters. To see a summary of the sampling, see Table 2.

	Headlines labeled as covering polarized topics	**Headlines labeled as not covering polarized topics**
Headlines from clusters labeled as covering polarized topics	78%	22%
Headlines from clusters labeled as not covering polarized topics	82%	18%

Table 2: Results from sample headlines

3.2 Method II (LDA) and Result Comparison

Next, we present the results from the second topic modeling method LDA. For the final model, we fit 10 topics through 10 epochs. For the model evaluation, we subjectively evaluated the readability of outputed topics. In both corpora, we were able to assign many topics at first glance. For instance Topic 2 in the Reddit World news corpus was interpreted as "Iran as a nuclear power" (see Figure 1, Appendix A). The remaining topics - although more difficult to label still seemed reasonable (see Tables 1 and 2, Appendix A).

With both methods, we observe that the Reddit corpus mostly covers global politics, including topics of immigration, trade, and nuclear weapons. On the other hand, the Ireland corpus contains a greater topic variety (see Figure 2, Appendix A and Table 2, Appendix B). In the results from Method II, this is expressed through most frequent words such as "album", "game", and "apple" reflecting respectively music, sports, and tech topics.

In terms of polarization, a similar manual labeling as in Method I again showed differences in polarization proportions between the datasets. From the LDA model, all ten topics from the Reddit corpus were identified as covering polarizing subjects, as opposed to less than half within the Irish Times headlines. This similarity allowed us to support the significance of the HDBSCAN results, on which we will build our final conclusion.

4 Discussion

The significance of our findings are mainly limited by the manual and thus subjective annotation of what topics are polarizing. Although we use a set definition for polarization, expert knowledge on the topic is missing. Furthermore, that a headline covers a polarizing topic does not mean that a headline itself is polarizing or leads to polarization. For future research, bringing in work on hate speech identification (Pérez-Escolar & Noguera-Vivo, 2021) or using sentiment analysis (Ali et al., 2021) could be helpful to distinguish "polarized" from "non-polarized" headlines or articles.

Moreover, during the HBDSCAN clustering, 43-46 % of headlines were left unlabeled. This could have the natural explanation that certain headlines contain very specific vocabulary which is represented by vectors with large euclidean distances from other points of data. However, it could also be due to the minimum cluster size used in the clustering. With a larger minimum size, smaller clusters might be counted as part of a larger cluster or left as noise. This exclusion of a large amount of headlines restricts the significance of our conclusion since much complexity of the data is lost. It should also be mentioned that headlines entirely made up by words missing from the word2vec model were completely filtered out during the embedding process. However, such headlines constituted less than 0.3% of each corpus so we assume that the data loss did not substantially impair our results. Nevertheless, it could reduce the perceived topic diversity in the datasets and minimize representation of niche topics in the clustering process.

Finally, our results align with the findings within media studies and previous qualitative analyses.We hypothesize that the difference in polarized topic coverage is a result of the differing purpose of Reddit and the Irish Times. Reddit is primarily a social media platform and exhibits characteristics of "click bait" (Blom & Hansen, 2015) and "sensationalism" (Arbaoui et al., 2016). On the other hand, the Irish Times is primarily a news source aiming to provide truthful and reliable information coverage. Contrasting topic polarization of this "traditional" newspaper with a tabloid could be interesting for future research.

5 Conclusion

Using both HDBSCAN clustering and Latent Dirichlet Allocation, we were able to demonstrate a significant difference in the coverage of polarized topics between the Reddit r/worldnews thread and the Irish Times headlines. Although the results of this study are limited by several factors, the empirical findings do align with the qualitative research conducted in the field of media studies.

6 Authors' Contribution

- Alice: Pre-processing pipeline, time-it function, abstract, datasets
- Dasha: LDA method code and corresponding report sections
- Fiammetta: Pre-processing code, README file, introduction, discussion, and conclusion
- Nora: HBDSCAN clustering method code, cluster labeling and corresponding report sections

References

Allaoui, M., Kherfi, M. & Cheriet, A. (2020). Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study. *Lecture Notes In Computer Science*. pp. 317-325.

Arbaoui, B., De Swert, K. & Brug, W. (2016). Sensationalism in News Coverage: A Comparative Study in 14 Television Systems. *Communication Research*. 47 pp. 299-320.

Azeem, A., Hunter, J. & Ruffman, T. (2019). News headlines or ideological beliefs: What affects readers' interpretations of news stories about immigration, killing in the name of religion and other topical issues? A cross-cultural analysis.. *New Zealand Journal Of Psychology*. 48 pp. 56-61.

Blei, D., Edu, B., Ng, A., Jordan, M. & Edu, J. (2003). Latent Dirichlet Allocation.. *Journal Of Machine Learning Research*. 3 pp. 993-1022.

Blom, J. & Hansen, K. (2015). Click bait: Forward-reference as lure in online news headlines. *Journal Of Pragmatics*. 76 pp. 87-100, https://www.sciencedirect.com/science/article/pii/S0378216614002410.

Campello, R., Moulavi, D. & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. *Advances In Knowledge Discovery And Data Mining*. pp. 160-172, https://link.springer.com/chapter/10.1007/978-3-642-37456-2_14.

Campello, R., Moulavi, D., Zimek, A. & Sander, J. (2015). Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. *ACM Transactions On Knowledge Discovery From Data*. 10 pp. 1-51.

Carothers, T. & O'Donohue, A. (2019). Democracies Divided: The Global Challenge of Political Polarization. Brookings Institution Press.

Chris. (2016). Worldnews on Reddit from 2008 to Today. Kaggle. https://www.kaggle.com/datasets/rootuser/worldnews-on-reddit

Google Archive. (2013). word2vec. Google. https://code.google.com/archive/p/word2vec/

Kroon, A., Trilling, D. & Raats, T. (2020). Guilty by Association: Using Word Embeddings to Measure Ethnic Stereotypes in News Coverage. *Journalism Mass Communication Quarterly*. https://doi.org/10.1177/1077699020932304

Kulkarni, R. (2022). Irish Times-Waxy-Wany News. Kaggle. https://www.kaggle.com/datasets/therohk/ireland-historical-news

Majid, A. (2022). Most popular websites for news in the world: Monthly top 50 listing. Press Gazette. https://pressgazette.co.uk/most-popular-websites-news-world-monthly/

Manikonda, L., Beigi, G., Liu, H. & Kambhampati, S. (2018). Twitter for Sparking a Movement, Reddit for Sharing the Moment: #metoo through the Lens of Social Media.

McCoy, J. & Somer, M. (2018). Toward a Theory of Pernicious Polarization and How It Harms Democracies: Comparative Evidence and Possible Remedies. *The ANNALS Of The American Academy Of Political And Social Science*. 681 pp. 234-271.

McInnes, L., Healy, J., Saul, N. & Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *Journal Of Open Source Software*. 3 pp. 861.

Moody, C. (2016). Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec.

Pérez-Escolar, M. & Noguera-Vivo, J. (2021). Hate Speech and Polarization in Participatory Society. Routledge.

Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal Of Computational And Applied Mathematics*. 20 pp. 53-65, https://ac.els-cdn.com/0377042787901257/1-s2.0-0377042787901257-main.pdf?_tid=6dc8e7be-5a14-4eb8-880c-2b477cb6d431acdnat=1552723814_f87e6e69a11ce3ad0892fa689c1da1a6

Rückle, E., Peyrard, S. & Gurevych, M. (2018). Concatenated Power Mean Word Embeddings as Universal Cross-Lingual Sentence Representations.

Wintrode, J. (2011). Using latent topic features to improve binary classification of spoken documents. *2011 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*.

Yan, R. & Gao, G. (2020). Topic Analysis by Exploring Headline Information. *Web Information Systems Engineering – WISE 2020*. pp. 129-142

Appendix A

Figure 1: Reddit World News Word Cloud Topics From the LDA Model

Figure 2: Ireland News Word Cloud Topics From the LDA Model

Table 1: Reddit World News Self-labeled Topics and Polarization From the LDA Model

Reddit World News	Topics	Polarizing
0	Political news (Russia, election, brexit)	1
1	Terrorism/war	1
2	Iran as a nuclear power	1
3	Syrian war	1
4	China & the Philippines maritime dispute	1
5	Turkey	1
6	Refugee	1
7	Terrorism	1
8	World news (Refugee, climate)	1
9	N. Korean nuclear missile launch	1

Table 2: Ireland News Self-labeled Topics and Polarization From the LDA Model

Ireland News	Topics	Polarizing
0	Government change	1
1	Life	0
2	Reviews	0
3	Court cases	0
4	World news	1
5	World news (Trump, Brexit)	1
6	Non-political news	0
7	Accident reports	0
8	Football	0
9	Terrorist attack in France	1

Appendix B

Table 1: Reddit World News Self-labeled Topics and Polarization From the HDBSCAN model

Label	Theme	Polarizing
-1	Unlabled Data	-
0	-	-
1	Summary of news	0
2	US involvement in international conflict	1
3	Information about terror organizations	0
4	War in middle east	1
5	War in middle east	1
6	Natural distaster	0
7	Nuclear power, Iran	1
8	Humanitarian Crisis	0
9	International crime and conflict	1
10	Internation diplomacy Israel-Palestine	1
11	-	-
12	-	-
13	Brexit	1
14	Big tech business	0
15	Information about terror organizations	1
16	Iran-Iraq conflict	1
17	Dead and missing people	0
18	Nuclear power, Iran	1
19	Stock market	0
20	Natural disaster	0
21	General economy	0
22	Humanitarian Crisis	0
23	Climate change	1
24	Climate change	1
25	Brexit	1
26	Food shortage	0
27	School Shooting	0

28	Culture	0
29	Tax system	1
30	US involvement in Israel-Palestine	1
31	Humanitarian Crisis	0
32	Natural distaster	0
33	Right to bear arms	1
34	Religious conflicts	1
35	Science	0
36	Iran-Iraq conflict	1
37	Nuclear power, Iran	1
38	Iran-Iraq conflict	1
39	Social welfare system	1
40	-	-
41	Violent interrogation (torture) by CIA	1
42	Conspiracy theories	1
43	Natural distaster	0
44	Natural distaster	0
45	Natural distaster	0
46	Crime, theft	0
47	Crime, theft	0
48	Animals	0
49	Drugs and substance use	1
50	American presidents	0
51	-	-
52	Nobel prize	0
53	US-Cuba conflict	1
54	Sports accomplishments	0
55	LGBTQAI+, gay marriage	1
56	Culture	0
57	Culture	0
58	Sports accomplishments	0
59	Epidemic	0

60	School Shooting	0
61	Accidents	0
62	Dead and missing people	0
63	Accidents	0
64	Crime, theft	0
65	Terrorist attack	0
66	Somali piracy	0
67	Religious conflicts	1
68	Epidemic	0
69	Epidemic	0
70	US-Cuba conflict	1
71	Internet in China	0
72	Accidents	0
73	Science	0
74	Culture	0
75	Culture	0
76	OG in Beijing, sports	0
77	OG in Beijing, boycott	1
78	China-Tibet conflict	1
79	Dead and missing people	0
80	OG in Beijing, boycott	1
81	OG in Beijing, boycott	1
82	Awards	0
83	Terrorist attack	0
84	Accidents	0
85	Sports	0
86	Terrorist attack	0
87	Natural distaster	0
88	Humanitarian Crisis	0
89	Humanitarian Crisis	0
90	Natural distaster	0
91	Natural distaster	0

92	International crime and conflict	1
93	Humanitarian Crisis	0
94	OG in Beijing, sports	0
95	Medical development	0
96	Natural distaster	0
97	Natural distaster	0
98	Humanitarian Crisis	0
99	-	-
100	Animals	0
101	Nature	0
102	International crime and conflict	1
103	Celebreties	0
104	Space	0
105	Dead and missing people	0
106	Sports	0
107	Fritzl case	0
108	Court cases, violence and homocide	0
109	US-Russia relations	1
110	Court cases, violence and homocide	0
111	Natural distaster	0
112	Celebreties	0
113	Culture	0
114	Elections	1
115	Elections	1
116	American presidents	0
117	National conflicts	1
118	Statistics	0
119	National conflicts	1
120	General economy	0
121	General economy	0
122	Tax system	1
123	Science	0

124	Terrorist attack	0
125	Stock market	0
126	General economy	0
127	US involvement in Iran-Iraq	1
128	US involvement in Iran-Iraq	1
129	US involvement in Iran-Iraq	1
130	International travel	0
131	Accidents	0
132	Animals	0
133	Viral videos	0
134	Dead and missing people	0
135	Nuclear power, Iran	1
136	Court cases, violence and homocide	0
137	Technological development	0
138	Accidents	0
139	Viral videos	0
140	Russian involvement in Balkan	1
141	Israel-Palestine conflict	1
142	Abortion rights	1
143	Religious conflicts	1
144	Court cases, violence and homocide	0
145	Court cases, violence and homocide	0
146	Iran-Iraq conflict	1
147	Afghanistan conflict	1
148	Nuclear weapons in Iran and NK	1
149	Dead and missing people	0
150	Low-profile court cases	0
151	International crime and conflict	1
152	Local wars	1
153	International crime and conflict	1
	Number of polarized clusters:	55

Table 2: Ireland News Self-labeled Topics and Polarization From the HDBSCAN model

Label	Theme	Polarizing
-1	Unlabeled data	-
0	-	-
1	-	-
2	-	-
3	-	-
4	Terrorist Attack	0
5	Brexit	1
6	Brexit	1
7	Culture	0
8	British Royalty	0
9	Irish local news	0
10	OG in Beijing, boycott	1
11	Iran-Iraq conflict	1
12	IRA	1
13	IRA	1
14	Stock market	0
15	Sports	0
16	Sports	0
17	Fishing	0
18	Natural Distaster	0
19	Natural Distaster	0
20	Sexual abuse of children	0
21	Economic growth	0
22	Humanitarian Crisis	0
23	Irish political conflict	1
24	Irish political conflict	1
25	Culture	0
26	Brexit	1
27	Fishing	0

28	Animals	0
29	International crime and conflict	1
30	Economy	0
31	Economy	0
32	Job market	0
33	Economy	0
34	Dead or missing people	0
35	Fritzl case	0
36	International crime and conflict	1
37	Dead or missing people	0
38	Elections	1
39	Technology	0
40	Internet	0
41	Culture	0
42	Science	0
43	Science	0
44	Dead or missing people	0
45	Brexit	1
46	Irish political conflict	1
47	Natural Distaster	0
48	Crime, robbery	0
49	Culture	0
50	People	0
51	Economy	0
52	Economic growth	0
53	Culture	0
54	General national politics	0
55	Economic scandal	0
56	Economic trends	0
57	Stock market	0
58	Stock market	0
59	Stock market	0

60	International politics	1
61	Tax system	1
62	Tax system	1
63	Dead or missing people	0
64	Accidents	0
65	Sports	0
66	Accidents	0
67	Dead or missing people	0
68	Culture	0
69	Accidents	0
70	Technology	0
71	Technology	0
72	Accidents	0
73	Dead or missing people	0
74	Irish history	0
75	Irish history	0
76	Culture	0
77	Irish culture	0
78	US presidential election	1
79	US presidential election	1
80	Culture	0
81	Memorials	0
82	Crime theft	0
83	Accidents	0
84	-	-
85	Tax system	1
86	Music	0
87	Epidemic	0
88	Epidemic	0
89	Crime, robbery	0
90	Job market	0
91	General Irish politics	0

92	General international politics	0
93	Accidents	0
94	Health	0
95	Accidents	0
96	Dead or missing people	0
97	Health	0
98	Terrorist Attack	0
99	Terrorist Attack	0
100	Terrorist Attack	0
101	International crime and conflict	1
102	Accidents	0
103	Accidents	0
104	Science	0
105	Awards	0
106	Culture	0
107	Awards	0
108	Irish involvement in international crime and conflict	1
109	Science	0
110	History	0
111	Dead or missing people	0
112	History	0
113	World War 2	0
114	International crime and conflict	1
115	Sports	0
116	Mental health	0
117	Culture	0
118	Culture	0
119	Culture	0
120	Culture	0
121	Culture	0
122	Culture	0
123	Terrorist Attack	0

124	Terrorist Attack	0
125	Irish history	0
126	General Irish politics	0
127	Sports	0
128	Sports	0
129	Sports	0
130	Sports	0
	Number of polarized clusters:	23