Computational Clustering of Media Biases on Indian Farmers' Protest Coverage: A Framework

Jagjot Singh
Msc. Data Science with Business
University of Exeter
Exeter, UK

Abstract—**The occurrence of an event when covered by news media can have large ramifications, particularly on how readers perceive the event. This puts a heavy burden on the shoulders of the media houses to inform the masses correctly and ethically in order to shape the public perception on general events. This power, when exercised in influence of biases, can lead the readers to absorb them. This study aims to use the latest NLP tools and techniques to identify the 'media biases' in news sources by analysis frames in the coverage of ongoing Indian farmers' protest. These protests are going on against the contentious farm laws introduced in the parliament in September 2020 to decentralise the farming infrastructure by making it conducive for private based enterprises. While studying this particular topic, this study aims to establish a framework with repeatable steps and low human interference to cluster multiple news frames and to form clusters of news sources that share the similar frames to identify biases in the news media**.

Index Terms—Media Frames, NLP, Unsupervised Learning

I. Introduction

With the rise in wide-spread access to the and in turn in the popularisation of social media platforms, the way people consume information has evolved in the past decades. People now have access to a large amount of information from a wide array of sources. And yet the studies have shown that even though people consume their news through different platforms, the general opinion of established newspapers, radio and TV news is of the more trustworthy and final (Guess et al., 2018, p 20). However, the news media does not take this trust for granted, especially when the trust in mass media is on a decline (Jones, 2004) and because it has to compete with multiple platforms for consumer attention. One of the ways they try to hold on the public attention is to focus on the dramatised aspects of the events (Beattie and Milojevich 2017). This gives us the opportunity to study different 'media frames' that news media covers about a certain topic and with it what they choose to write about and ignore completely. Studying this can be used to infer biases in the news media and help make a journalistic tool to evaluate the quality of the news consumed.

The concept of frame is studied in multiple domains and has no unanimous definition which is agreed upon by the researchers. This was recognised as a 'fractured paradigm' by Robert Entman (1993) and he went ahead to define framing as: *'to frame is to select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described'* (Entman, 1993). This definition is widely taken as a pivot point in many research papers that influenced this report (Burscher et al., 2014; Shurafa et al. , 2020; Walter & Ophir, 2019; Rashed et al., 2020; Smith et al., 2020). In terms of news media, a frame can be inferred as the extracted salience of an event by exploiting the images, metaphors, stereotypes related to it. In this study we will try to extract frames used by the news media through headlines of the articles published on the coverage of Indian farmers' protest during the course of 8 months of its peak popularity. The analysis will try to group the news sources. The news media data is extracted from a news repository called Media Cloud, it is a consortium research project shared by University of Massachusetts Amherst, Northeastern University, and Harvard University. They collect and group news media data and related social media data from the news sources from multiple countries to understand the news media coverage (Díaz-Sánchez, et al., 2011).

This study will try to find media biases by extracting the frames from the headlines by using mBert sentence transformer, grouping the newspapers using the similar frames together by reducing the density based cluster algorithm, and then discovering the news sources that are clustered together throughout the course of the protest coverage by analysing the weighted graphs generated from them. The aim of this study is to establish this practice as a standard framework and hence, the research questions this study will delve into are:

RQ1: Could the latest sentence transformers be used to extract frames from short text ie. the headlines?

RQ2: Could the multilingual text, sharing similar context, be clustered together?

RQ3: And could the insights from the headlines of the newspapers in the same clusters be used to determine their inherent biases?

In the next sections, this report will briefly elaborate the back on the farmers' protests and motivation behind the study, and will also contextualise the gaps in the research on the subject (Section II). In section III, the report will give a brief introduction to the data source and the data used before explaining the experiment design and the decisions that led to it in section IV. In addition, this report will present the result of the experimentations (section V), followed by the discussion and possible future work (Section VI). Finally, we will briefly draw conclusions (section VII).

II. CONTEXTUALISATION & MOTIVATION

In this section, we explain the motivation behind the research by the means of contextualising the ongoing protests in India and literature review on news media framing analysis.

A. Indian Farmers' Protest

The Indian constitution divides the responsibility of agriculture management largely with the state governments while the central government is responsible for governing inter-state and international trade on agricultural commodities, states manage the sale of the produce through a state level regulatory body called Agricultural Produce Marketing Committee (APMC) market or *mandi (locally)*. APMC provides an auction market for farmers to sell their produce to the traders and intermediaries to whom they also issue licenses to buy. The auction begins at a Minimum Selling Price (MSP) which is set by the central government for 26 out of 70 crops produced in India which cover 85% of production[1]. In September 2020 the government introduced the three farm bills to change this infrastructure, by encouraging private players to establish a parallel central auction market against APMC where farmers can sell their produce, they gained assent from the President on 27th September 2020 and started the current protest. These laws are:

- **The Farmer's Produce Trade & Commerce Act, 2020**[2], limits State's control on APMC, while also allowing trade of produce to transact on a central electronic trade platform.
- **Essential Commodities Act (Amendment), 2020**[3]**,** pulls out some of the restrictions regarding hoarding and market manipulation that were on private businesses from an already existing law.
- **Farmers' agreement on Price Assurance and Price Services Act, 2020**[4], allows private sponsors to engage in written contracts with the farmers which are outside the purview of the 'State Act' and APMC. It also establishes a three-level settlement system, namely the conciliation board, Sub-Divisional Magistrate and an Appellate Authority to resolve any disputes among both parties and to uphold the said contract.

Farmer unions found these laws skewed towards the private organisations and against the farmers. They criticised the first act, which introduces a central trade market, and remarked that it does not mention Minimum Support Price (MSP) and that corporations with access to better lawyers can take advantage of this caveat against farmers with considerably lesser resources. They also criticised the regulation under the third law, which limits the highest level appeal for the farmers to an Appellate Authority and thus prevent them from going to the courts in case they feel wronged. (Narayanan., 2020)

On November 25, 2020, farmers started to congregate in masses to protest against the three laws around the Northern and Western border of the National Capital Region of New Delhi. This congregation was met by resistance by the local police. After multiple rounds of inconclusive meetings farmer unions decided to carry out a 'tractor rally' towards the capital city on the Republic Day (25th January). This procession led to a violent and chaotic rioting during which public transport was vandalised and 'farmer movement' flags were hoisted on the Red Fort, New Delhi. However, the protesters are still gathered at the borders of the NCR at the time of writing this report but the news coverage on the topic has substantially decreased.

The major events of the protests are widely reported by all major news sources in India, reports included advantages of the reforms, cause of the protest, police brutality, violent protest, locals supporting the protesters etc. While it is difficult to judge whether either party is on the right side of the argument, the coverage of the events can shed light on what frames were used by the news media during the course of 8 months. This gives us the opportunity to study the coverage of the various events and to analyse how often some news media sources used similar headlines and covered similar events. Also, an analysis on the frames from multiple sources allows us to group media sources, insight to which could be used to infer biases in them.

[1] https://sites.ndtv.com/cultivatinghope/project/minimum-support-price-harsh-reality-vs-good-intent/

[2] http://egazette.nic.in/WriteReadData/2020/222039.pdf

[3] http://egazette.nic.in/WriteReadData/2020/222038.pdf

[4] http://egazette.nic.in/WriteReadData/2020/222038.pdf

B. Literature Review

The definition of framing stated by Entman in 1993 covers a very broad spectrum of possibilities, since then the framing theory has evolved and frames are attempted to be classified into narrower definitions based on the application and context. Studies pertaining to news media frames in terms information perception segregate frames as, *equivalence frames*, which is presenting the same information but stating them differently (Chong & Druckman, 2007a), and *emphasis frame,* which focuses on the certain aspect of the information (Chong & Druckman, 2007b). In the context theme of the issue being discussed, news media frames have been defined such as, *issue-specific* frames and *generic* frames (De Vreese, 2005). We are agnostic about an absolute definition or classification to frame, because while trying to compute frames about the coverage of a topic it is confusing to label a group of headlines to issue specific frame or emphasis frame, as it seems to be applicable to both the definitions. However, in a number of studies to computationally find media frames, regardless of the confusion with labelling an established type, researchers tend to define their frames depending on the result they wish to produce or the data being analysed.

The difference in different frames can sometimes be very subtle structurally for computational methods to label them uniquely and hence, choice of data can make a huge difference in the quality of the analysis and definition of frame. Analysing twitter data can be comparatively easy, given that hashtag analysis can be a straightforward technique to separate possibly opposing narratives (Shurafa et al. , 2020). Topic modelling (LDA) has been used to describe news frames for unlabeled data, but it has been warned as unreliable and is usually paired with other methods to produce quality results as it tends to group opposing sentiments because they used the same phrases, which can lead to internally incoherent and externally similar frames (Smith et al. 2020). One such example of it is topics using topic modelling (LDA) on the labelled news data from Lexis-Nexis database on US election coverage which is labelled according to candidates about whom they were published (Walter & Ophir, 2019). This research uses network analysis to connect different frames in communities of topics discovered per candidate. Labelled data, however, are few in number and of those available very rarely open-source. Hence, researchers also manually label their data to build models that can identify particular frames which suit their research purpose. Such as defining frames as biases (personalisation, dramatisation and negativity bias), to analyse how the biases in coverage on a particular topic has changed over the years (Opperhuizen et al., 2019) or defining frames by scoring devices such as strength, power and sentiment used for the accused and accusing parties in the coverage of #MeToo movement (Field, Bhat & Tsvetkov, 2019). Machine learning models using such data are highly accurate in defining frames, but their scope is limited to the frames which are defined by the coders. Also, because framing is greatly dependent on the individual perception of the text, the frames in coder labelled data have also been found to be different for the same text (Burscher et al., 2014). Another approach, which influenced the choice of method in this project, leverages the merits of word embedding to identify biases (Caliskan, Bryson, and Narayanan, 2017) and to use labelled data to measure its potential. For instance, deep learning transformers such as BERT have been used to successfully cluster the frames in the tweets of the user supporting or opposing candidates of a political party (Rashed et al., 2020). Tweets, like headlines, are short text and in this study we exploit word embedding models to cluster similar headlines to find biases. We define frames as the presence or absence in reporting of a certain event and aim to cluster, BERT, contextually similar headlines.

C. Motivation

Frames that the media use form our perception on the topics being covered and can have serious consequences. Nelson, Oxley & Clawson (1997) presented evidence of such an effect on Ku Klux Klan (KKK) rallies, where people following free-speech frames were more tolerant towards KKK than the people who followed public order issue frames. This study emphasises the role of the Media as the 5th pillar of democracy and the gravity of the responsibility that rests on their shoulders.

Media, however, like any other business is answerable to their shareholders and wants a wider audience, which leads them to present news in a certain way that appeases their target audience based on their general and political stance (Bennet, 2009). It is hence argued, not only the media sources can be biased but it is more profitable for them to be so. An effect of which is polarisation and difficulty to agree on tenuous topics (Sunstein, 2002).

The computational media framing tools are yet to advance and applications of the tools are limited or dependent on manual labelling and can only give insight into a specific topic. This study is motivated by this challenge and aims to set a foundation in developing a journalistic tool which can use the unsupervised methods without manually labelling and to induce coder's biases to the results while effectively differentiating the biased news sources on any topic.

This study also tries to build a framework with repeatable steps that can be applied to the headlines of any topic and to group together the news sources which covered similar events throughout the course of the topic covered.

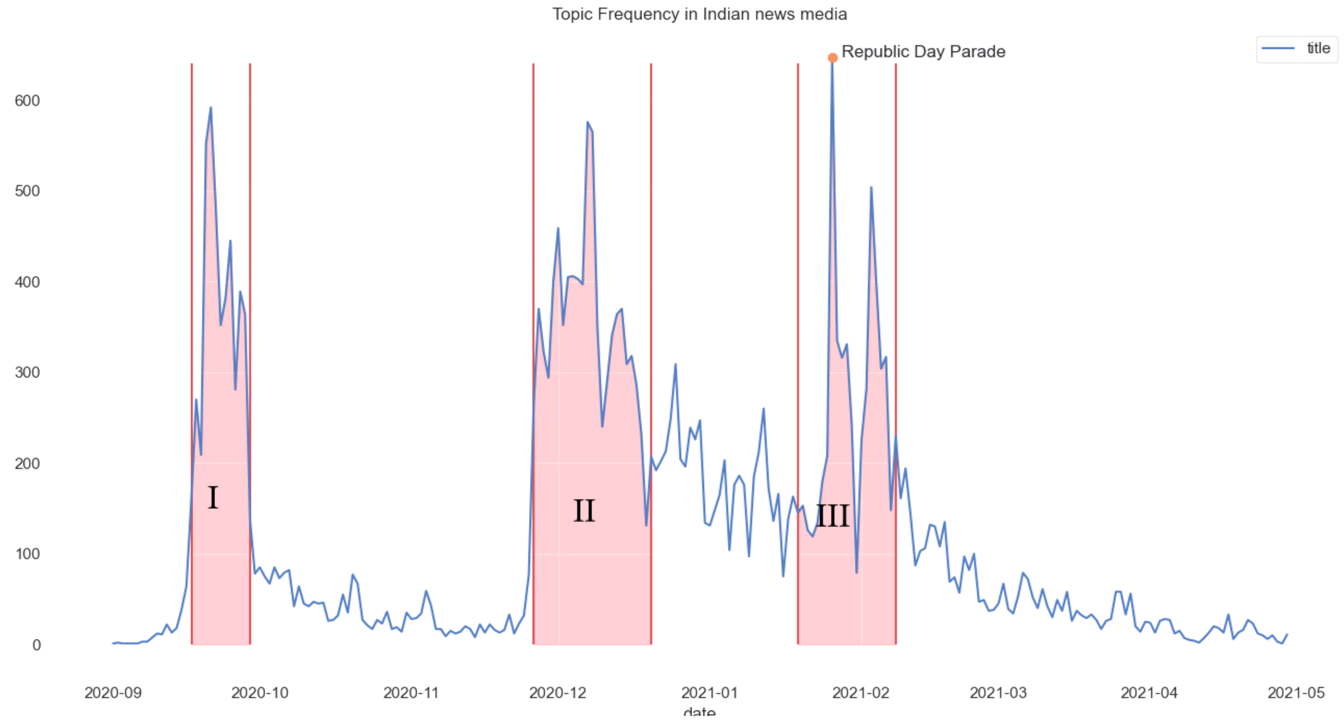


Figure 1 - News article frequency for the selected timeline. Groups 1, 2 and 3 are highlighted in red.

III. UNDERSTANDING THE DATA

Multiple factors weighed in with our selection of **Media Cloud**[5] as the news media source for our analysis. Firstly, it has a collection of recent Indian news topics from over 195 India news sources, including the headline, dates and URL to the story. In addition, unlike Lexis-Nexis it is not limited to the labelled data available and unlike it, has an open source API connection which is accessible free of charge. Lastly, the Media Cloud does ample pre-processing to group the news titles together which are covering the similar topic, from there the users can either search all the coverage using valid keywords on the UI of their site or use the query using an open-access API. We used the former method by using ‘farmers protest’ and ‘farm laws’ as keywords among Indian news sources. By following the number of articles written on the topic we chose the timeline between September 1st, 2020 and April 30th, 2021.

Throughout the analysis we take several actions to narrow down the data to make it more specific and improve accuracy of the analysis. Initially the total number of headlines we procured in the keyword search and filtering out entries with missing date values was 29931. Then we chose the top 36 news sources which have covered 90% of the total news based on the numbers of articles per source, bringing the number of articles in consideration to **27069**.

The number of news articles published per day show clearly that there are a few peaks and the data is divided into certain groups. These peaks also correspond to real life events and guided us to analyse them separately. These separations were made to ensure that the headlines published by the news sources fall close to each other when embedding is done (see Section IV) so groups which are made in the process are more accurate. Segregating the data also defines three event groups and provides a basis to find the newspapers which are always clustered together throughout the three groups. Figure 1 shows the frequency of the news articles published per day for the topic, and the aforementioned groups are shaded in red. The events which correspond to the groups are (Figure 1):

- **Group 1** *(17th Sept 2020 - 29th Sept 2020)***:**
The laws were approved in the houses of the parliament, Lok Sabha on 17th September and in Rajya Sabha on 20th September 2020. This sparked local protests. The President gave assent to the laws on 27th September 2020.
(4615 articles)

[5] https://mediacloud.org/

- **Group 2** *(26th Nov 2020 - 20th Dec 2020)***:**
The farmers started to congregate around the capital city and met resistance by local police authorities on 26th November. Throughout the month, farmers had talks with government officials and filed a complaint to the Supreme Court against the laws on 11th December and apex court suggested the government to hold the laws on 16th December 2020. During this time frame the topic was most popular and had become national news.
(8654 articles)

- **Group 3** *(19th Jan 2021 - 8th Feb 2021)***:**
Farmers planned a massive rally coinciding with the Republic Day of India on 26th January. On 24th January, police granted the rally a fixed route. On the day of the rally, the protesters moved away from the fixed route and entered the central part of the city. In the coming days police attempted to clear the borders where farmers were protesting.
(5426 articles)

The segregation of the groups should reflect in the headlines of the data as well. It is important to consider that we are using headlines as our primary data and because it is short text, topic modelling (LDA) does not work particularly well on such data (Jónsson, E., & Stolee, J. (2015)). Instead, to display the difference between the three groups we used TF-IDF ranking to find the most relevant words used in different groups to validate the selection of the groups. This requires pre-processing of the data, details of which are elaborated in the next section. The next section will also put forth the proposed framework to generalise the process of grouping the newspapers.

III. METHOD : THE FRAMEWORK

The steps of the process are also referred to as a 'framework' in the report because the process defines a number of steps that can be recreated to find biases in the news sources for any topic which is widely covered and headlines are sourced from Media Cloud platform.

The proposed framework aims to be as unsupervised as possible so human intervention should not induce any biases in the process. Hence, the parts where human intervention is suggested have taken into account that they only improve the accuracy of the process without compromising the process, such as dividing the topic into groups based on the events. Dividing the events into groups requires the knowledge of the timeline of the events, and doing so provides the ability to compare the clusters in different groups. However, to validate that the selection covers different events in the topic we use TF-IDF ranking of the most relevant words. The steps of the framework are as mentioned below:

A. Generating Lemmas

Another step where human intervention may be required is to create lemmas from the synonymously used words (often nouns) or multiple words used together, so that they are taken as one when processed by NLP algorithms. This step is optional and requires context of the words that are being used. The pre-processing rules for the data used are shown in Table 1.

Words	**Lemmas/Substitutes**
Farm Laws	FarmLaws
Farmers' Protests	FarmersProtest
Oppn/INC/UPA	Congress
Bharatiya Janata Party	BJP
Shiromani Akaali Dal/ Akaali Dal/ Akaais/ SAD	AkaliDal
Prime Minister/Modi/ Narendra Modi	PMMODI
RS/Rajya Sabha	RajyaSabha
Farmer Union	FarmerUnion
Farm Bills	FarmBills

Table 1 - Pre-processing: Combining words that are used together or synonymously into single words/lemmas.

B. TF-IDF ranking

TF-IDF or Term Frequency - Inverse Document Frequency is a statistical method used to find the most relevant words. It is done by multiplication of frequency of term in the document (or sentences) and inverse of the log of the word across the set of documents (or sentences).

This is a validation step for the selection of the aforementioned groups (Section III). The most relevant words should be different for each group and should reflect the basis on which the groups are selected, if that is not the case it is suggested that the groups are shuffled again. Furthermore, comparing the TF-IDF ranking scores for each news source can give a gist of how different news sources have covered the events in the timeline of each group.

For the topic under the purview of this report, the most relevant words according to TF-IDF validate the selection of the groups (Table 2). Group 1 highlights the 'Farm Bill' and 'Punjab' is mentioned as well as 'protests', this indicates that the protests were regional, unlike to that of 'Group 2' which mentions 'Delhi' and 'border', it also mentions 'Farmers Protest' as an important word which indicates that these words were often used together in the newspapers. Group 3, however, mentions words such as 'violence', 'tractor', 'rally' and 'republic'. In conclusion, the table indicates that the basis of selecting the groups stands valid (See also, Appendix 1).

TF-IDF Ranking	**Group 1**	**Group 2**	**Group 3**
1	bills	farmers	farmers
2	farm	protest	protest
3	farmers	delhi	delhi
4	congress	protesting	police
5	rajyasabha	farm	tractor
6	farmbills	laws	day
7	protest	border	farm
8	pmmodi	centre	violence
9	punjab	pmmodi	rally
10	bjp	says	laws
11	says	govt	पर
12	mps	bandh	says
13	protests	support	protesting
14	parliament	bharat	pmmodi
15	govt	farmersprotest	republic

Table 2 - Top 15 most relevant words according to TF-IDF rankings

C. Word Embedding using mBERT Transformer

For the purpose of clustering of headline text it was required to vectorise the text in a way that the vectors of contextually similar headlines are closer to each other. We used the mBERT transformer to vectorise each of the headlines in the groups. BERT (Bi-directional Encoder Representation from Transformers) is a bidirectional transformer which uses an attention mechanism to encode the sentences by reading them both left-to-right and right-to-left. BERT is pre-trained on unlabelled text from wikipedia and book corpus dataset, which is primarily in English language and mBERT (multilingual BERT), on the other hand, works similarly but is trained on 104 languages where languages with lower volume of data is scaled. It was made by Google for predictive text writing (Devlin et al., 2019). It works well with smaller text, and in fact has the limit of 512 tokens at a time. Hence we used multilingual BERT for word embedding for headlines in our data.

D. Dimensionality Reduction and Clustering

mBert converted our text into an encoded vector of high dimensionality with 768 values, which places all the vectors close to each other and makes it difficult to extract meaningful clusters (Fan & Li, 2006). UMAP (Uniform Manifold Approximation and Projection) is a dimensionality reduction technique which conserves the global and local structure of the high dimensionality data by approximating the clusters by making a graph where edges are weighed by using k nearest neighbours of each point and then projecting them closely or farther from each other in a lower dimension based on probability of those edges (McInnes et al., 2018).

For clustering the UMAP projections we used HDBSCAN algorithm, which gives out density based stable cluster by transforming points according to distance, connecting the closer points and making a minimum spanning tree based on them (Campello et al., 2013). This method work similar to DBSCAN in principle and tries to improve upon it by uses h

The output from both of these methods depend highly on hyperparameters used for them, we used optimised the hyperparameters by using the number of outliers classified as our metric. We used these steps for each of the groups and the whole text (Appendix 2). Former process resulted in tighter groups and lower number of outliers overall. We were able to identify more than 50 clusters for each of the groups and topics were grouped according to the number similarity in the text and the language used, for example, Table 3 one such randomly selected cluster from group 1 (Figure 2) which has contextually similar headlines in Hindi and in English from clusters. Each cluster is assumed to be a distinct frame. For lucidity, we will call the result, 'the headline clusters'.

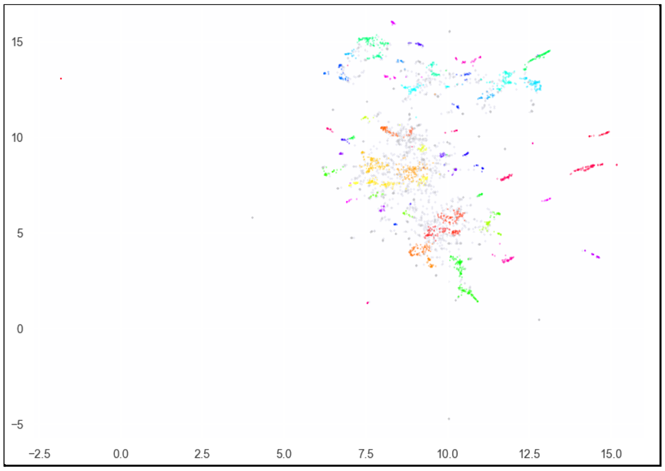


Figure 2 - 2-D representation of clustering results for Group 1 after mBERT, UMAP and HDBSCAN

Headlines from a label from Group 1
President nod To Farm Bills: Capt explores state law tweak; Sukhbir says ready to join any struggle',
Amid intensifying protests, President Kovind gives assent to controversial farm bills, laws come into force immediately',
President Kovind gives assent to 3 farm bills passed by Parliament',
"President's nod to farm bills 'sad, disappointing and extremely unfortunate', says Sukhbir Singh Badal",
"'Extremely unfortunate': Sukhbir Singh Badal on Presidential assent to farmers and J-K bills",
President Kovind gives assent to farm bills amid protests; Maharashtra says no to implementation',
President Ram Nath Kovind Gives Assent To Contentious Farm Bills',
"'Sad, disappointing': Sukhbir as farm Bills get President's nod",
राष्ट्रपति रामनाथ कोविंद ने तीनों संशोधित कृषि बिल पर हस्ताक्षर किए',
"'Sad, Disappointing': Sukhbir Singh Badal as Farm Bills Get President Kovind's Approval",
"'Sad', says Sukhbir as farm bills get Prez nod",

Table 3 - Randomly selected cluster with similar context headlines from Group 1. (The headline cluster)

F. Generating a Weighted Network and Clustering

We use the clusters of contextually similar headlines or frames. We can extract the newspapers using similar frames over the course of the protest. We use a weighted network where nodes are represented by the news sources and links are calculated using the clusters. We used multiple approaches to use the clusters from the previous sections to represent a generalised weighted edge. By 'generalised weighted edge' we mean that edge measure that can represent faithfully all types of data and takes into consideration different frequency of news sources which have widely covered the topic and of those which have scarcely covered it.

We did this by calculating the proportion of the headlines each newspaper has in each cluster, making a vector from the newspapers with each value representing the proportion of the news covered in all different clusters sequentially. Using proportions instead of counts made sure that high weights do not favour the highly frequent news sources. Secondly, we calculate cosine similarity between each of these news sources to get a measure of weight which will represent the connection of the nodes in the graph. We did not use correlation as our method to calculate similarity so that we only get values of the weights between 0 to 1, which makes it easier for node clustering algorithms to find tight clusters. Table 4 shows some of the news sources which have high similarity. We will use these similarities as weights for the network clustering in the next section.

Source A	**Source B**	**Wt. G1**	**Wt. G2**	**Wt. 3**
The Times of India	HindustanTimes	0.858	0.871	0.489
The Times of India	Indian Express	0.847	0.767	0.821
newindianexpress	sify.com	0.825	0.853	0.841
The Times of India	News 18	0.820	0.832	0.584
The Times of India	sify.com	0.797	0.837	0.864

Table 4 - Highly similar news sources based on weights calculated using cosine similarity vectors made by proportion of news sources appear in the same clusters.

G. Finding Tightly Clustered News Sources

We create a graph using the weights for each group of events mentioned above. We find newspapers which are tightly grouped together using the multilevel modularity maximization method suggested by Blondel et al. (2008); also referred to as “Louvain”. Comparing the results from the clusters from all three graphs we discover the news sources which have shared the same cluster throughout the timeline of the event.

We validate this approach by using a third party application specialised in visualising graphs, called ‘Gephi’. Using it we appended all the graphs over each other and used an in-built modularity function (Louvain) to identify clusters.

The resultant clusters turned out to be similar (if not exact, because of different random states in the Louvain method) to those which we calculated using our experiments (Figure 3). Finally, the results we present to the users of this proposed framework are the news sources which are clustered together in the weighted network generated. We call the final result, ‘the newspaper clusters’. The analysis of these news sources can be done by finding the headlines which these news sources share among each other. With this step, we will conclude the implementation of the proposed framework and with the result provided, users shall extract the biases, if any, in the news sources through inferring the nature of continuous frames used by the news sources.

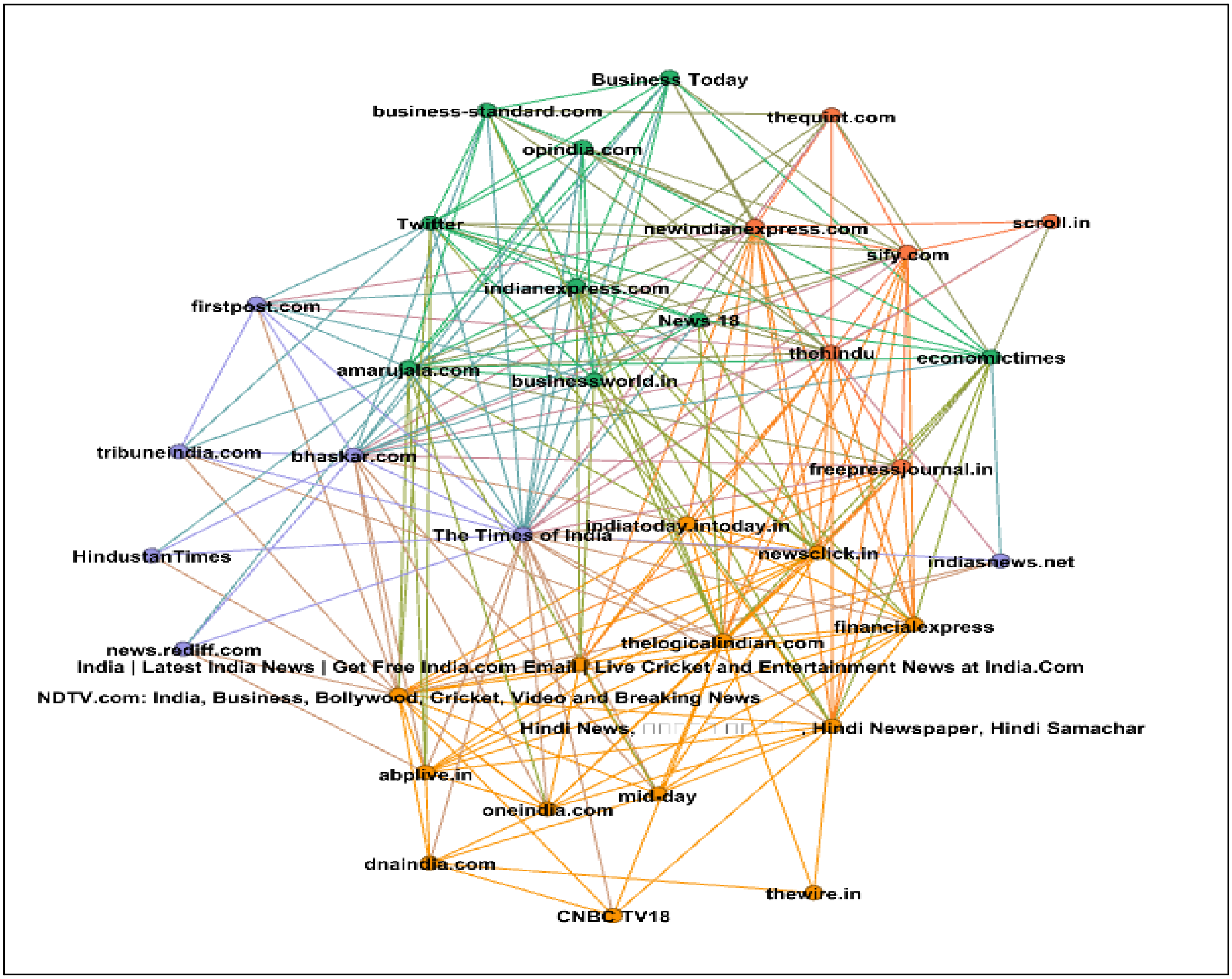


Figure 3 - Graph prepared appending three graphs on gephi. Clusters are found using Louvain with resolution 1. (The Newspaper Clusters)

V. Results

The results produced present the news sources which have covered the similar topics or used the same frames throughout the timeline of the topic covered. For the topic of Indian Farmers' Protest for the time of 8 months, the proposed framework produced the result as shown in Table 5.

#	Grouped News sources
1	['businessworld.in', 'News 18', 'indianexpress.com', 'firstpost.com']
2	['thehindu', 'sify.com', 'indiasnews.net', 'newindianexpress.com', 'Business Today', 'thequint.com', 'freepressjournal.in', 'opindia.com', 'scroll.in', 'economictimes', 'business-standard.com'],
3	['NDTV.com', 'abplive.in', 'thewire.in', 'oneindia.com', 'India.Com', 'indiatoday.intoday.in', 'CNBC TV18', 'thelogicalindian.com', 'dnaindia.com', 'news.rediff.com', 'financialexpress', 'newsclick.in', 'Hindi News', 'mid-day'],
4	['amarujala.com','The Times of India', 'HindustanTimes', 'bhaskar.com', 'tribuneindia.com']]

Table 5 - Group of news sources which have used similar language to report events throughout the protest. (Similar to the results in Figure3)

The framework is aimed to be a representation of grouped news sources which have reported events in the same way or have covered the same event most often together so that the analysts or journalists using the framework can analyse how the newspapers have covered the topics. This can be done in many ways but we chose to find the most relevant words using the TF-IDF algorithm again. We use the algorithm on the most popular 'headline clusters' used by the group of newspapers clustered together. The most popular cluster is extracted from the headlines clustered together after part D of section III. TF-IDF, however, is not the ideal tool to find the most relevant words because it separately scores Hindi words which are synonymous to highly scored English words, still it will give a clear indication of how a topic is covered. For instance TF-IDF scores for each group (1,2, and 3) of events for the most common cluster of headlines for the fourth newspaper group (Amar Ujala, The Times of India, Hindustan Times, Bhaskar and Tribune) in Table 5 is that for the group 1 of the time-series they have focused more on the road-blocks and rail-blocks by the protesters, then for group 2, the news sources have covered the clash between police and protesters, and also on the camp that was set-up by farmers on the borders of the capital, talking about medical and food facilities, as can be inferred from Table 6. Lastly, for the events transpired in group 3, the news sources have covered road-blocks again. The focus of these sources seems to be on interruptions which happened during the protest. These observations can be used by the journalists or an SME for the topics to infer if which topics have not been covered widely.

TF-IDF Ranking	groupOne	groupTwo	groupThree
1	farmers	farmers	chakka
2	bills	border	jam
3	farm	singhu	farmers
4	block	delhi	ईव
5	protest	protest	roads
6	rail	protesting	block
7	roads	security	शनल
8	traffic	police	पर
9	hit	पर	haryana
10	शन	doctors	punjab
11	रदर	tikri	कर
12	noida	camp	और
13	haryana	medical	expressway
14	यक	सप	top
15	highway	food	february

Table 6 - Most relevant words used in the corpus of the all the news source grouped together (in Amar Ujala, The Times of India, Hindustan Times, Bhaskar and Tribune)

VI. DISCUSSION AND FUTURE WORK

The aim of the project was to create a framework with repeatable steps to infer biases in the news coverage of a topic. It was designed keeping in mind that an evolved version of this framework can serve as a journalistic tool to analyse how the topics are covered in news media using the frame analysis. An evolved version could use the benefit of the open API feature from media cloud that the tool also becomes the portal to select the event and date range. On the other hand, the groups that we assume as frames, however, have scope of improvement, because the method groups headlines which are talking about the same issue, but misses to separate the clusters based on how these topics are being talked about. It is apparent in the results shown in Table 3, where some headlines talk about the bills as 'contentious', some talk about how a certain Chief Minister is disappointed by the President's assent and some factually state that the bills have been passed. Perhaps, a similar approach on day wise sets can be applied to separate these frames into smaller groups. Additionally, experimentation with other multilingual transformers such as BERT trained in Indian languages such as IndicBERT, or a more latest transformer such as XML-R (Conneau et al., 2019) can be done and their results analysed. Meanwhile, this approach is limited to what news sources chose to cover from a certain topic, not how they chose to cover it.

There is another challenge to this approach is variable hyperparameters for UMAP and HDBSCAN can change the results entirely. The present approach selects hyperparameters based on the lowest numbers of outliers to allow most numbers of headlines to be grouped. The problem with this approach is that the result is hard to validate any other way but manually. We tried the LDA approach to find the topics being talked about in each cluster to see how they separate from each other, but as mentioned earlier it does not work well with short text and it also would not work on multilingual data. Perhaps, a corrective measure could be to use bitwise-LDA (Liu & Chen, 2017) with translated data. Unfortunately, for either approach we do not have a stable python library and these would have to be developed internally.

Another clustering approach, which can be unreliable, is multilevel modularity maximising method. The results from this method change slightly each time because of inherent randomness in the algorithm. The risk that it carries is that it gives the user of the framework the 'newspaper clusters' which have some edging sources which can either be in the present or neighbouring group depending on the random state.This issue is apparent in the difference between the clusters in Figure 3, which shows the clusters from 3rd party software Gephi and Table 5, which are found by finding intersection of newspaper groups for three event groups.

VII. CONCLUSION

In conclusion, this project takes a novel unsupervised approach using a combination of state-of-the-art methods to compute biases in the news coverage. The framework and its limitations can become a solid foundation to improve upon and build a journalistic tool to monitor media coverage. Despite not having been able to answer the 2 out of 3 research questions that were laid out in section I of the project, this project does find that it is possible to cluster headlines sharing the same context, for RQ2. The project also partially answers RQ3, we were able to cluster news sources using the same headlines throughout, but the results are not always the same owing to the randomness in network clustering methods. As for finding frames using the transformers which are more insightful than just finding out what topics are covered not how they are covered, more work needs to be done on our approach for it to be able to recognise frames and for us to be able to validate them.

DECLARATIONS

Declaration of Originality.

I am aware of and understand the University of Exeter's policy on plagiarism and I certify that this assignment is my own work, except where indicated by referencing, and that I have followed the good academic practices.

Declaration of Ethical Concerns.

This work does not raise any ethical issues. No human or animal subjects are involved neither has personal data of human subjects been processed. Also no security or safety critical activities have been carried out.

REFERENCES

Guess, Andrew, Brendan Nyhan, and Jason Reifler. "All Media Trust Is Local." Findings from the (2018).

Jones, D. A. (2004). Why Americans don’t trust the media: A preliminary analysis.Harvard International Journal of Press/Politics,9(2), 60–75. https://doi.org/10.1177/1081180X04263461

Beattie, Peter, and Jovan Milojevich. 2017. “A Test of the ‘News Diversity’ Standard.” The International Journal of Press/Politics 22 (1): 3–22

Entman, Robert M. 1993. “Framing: Toward Clarification of a Fractured Paradigm.” Journal of Communication 43 (4): 51–58

Burscher, B., Odijk, D., Vliegenthart, R., Rijke, M. D., & de Vreese, C. H. (2014). Teaching the computer to code frames in news: Comparing two supervised machine learning approaches to frame analysis. Communication Methods and Measures, 8(3), 190–206.

Díaz-Sánchez, D., Almenarez, F., Marín, A., Proserpio, D., & Cabarcos, P. A. (2011). Media cloud: an open cloud computing middleware for content management. IEEE Transactions on Consumer Electronics, 57(2), 970-978.

Narayanan, S. (2020). Understanding Farmer Protests in India. Academics Stand Against Poverty, 1(1).

Chong , D., & Druckman , J. N. (2007a). A theory of framing and opinion formation in competitive elite environments . Journal of Communication , 57 , 99 – 118 .

Chong , D., & Druckman , J. N. (2007b). Framing theory. Annual Review of Political Science , 10, 103 – 126 .

De Vreese, C. H. (2005). News framing: Theory and typology. Information design journal & document design, 13(1).

Shurafa, C., Darwish, K., & Zaghouani, W. (2020, October). Political Framing: US COVID19 Blame Game. In International Conference on Social Informatics (pp. 333-351). Springer, Cham.

Walter, D., & Ophir, Y. (2019). News frame analysis: An inductive mixed-method computational approach. Communication Methods and Measures, 13(4), 248-266.

Smith, A., Tofu, D. A., Jalal, M., Halim, E. E., Sun, Y., Akavoor, V., ... & Wijaya, D. (2020). OpenFraming: We brought the ML; you bring the data. Interact with your data and discover its frames. arXiv preprint arXiv:2008.06974.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues.

In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Association for Computational Linguistics, Beijing, China, 438–444. https://doi.org/10.3115/v1/P15-2072

Opperhuizen, A. E., Schouten, K., & Klijn, E. H. (2019). Framing a conflict! How media report on earthquake risks caused by gas drilling: A longitudinal analysis using machine learning techniques of media reporting on gas drilling from 1990 to 2015. Journalism Studies, 20(5), 714-734.

Field, A., Bhat, G., & Tsvetkov, Y. (2019, July). Contextual affective analysis: A case study of people portrayals in online# metoo stories. In Proceedings of the international AAAI conference on web and social media (Vol. 13, pp. 158-169).

Burscher, B., Odijk, D., Vliegenthart, R., Rijke, M. D., & de Vreese, C. H. (2014). Teaching the computer to code frames in news: Comparing two supervised machine learning approaches to frame analysis. Communication Methods and Measures, 8(3), 190–206. https://doi.org/10.1080/19312458.2014.937527

Rashed, A., Kutlu, M., Darwish, K., Elsayed, T., & Bayrak, C. (2020). Embeddings-Based Clustering for Target Specific Stances: The Case of a Polarized Turkey. arXiv preprint arXiv:2005.09649.

Bennett, W. Lance. 2009. News. The Politics of Illusion. 8th ed. New York: Pearson Longman

Sunstein, C. R. (2002). Risk and reason: Safety, law, and the environment. Cambridge University Press.

Jónsson, E., & Stolee, J. (2015). An evaluation of topic modelling techniques for twitter. University of Toronto.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Fan, J., & Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. arXiv preprint math/0602133.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

Campello, R. J., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In Pacific-Asia conference on knowledge discovery and data mining (pp. 160-172). Springer, Berlin, Heidelberg.

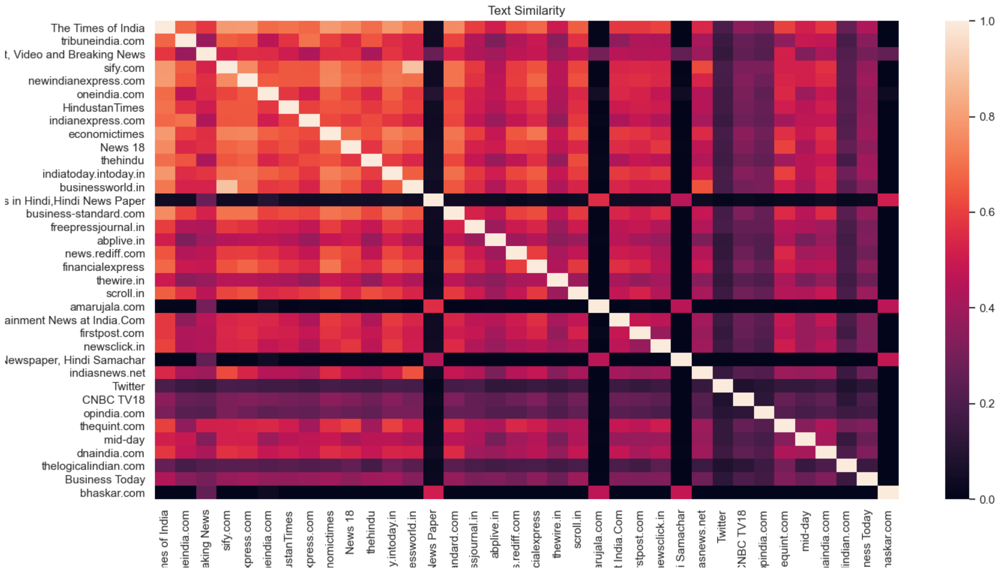
Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 2008(10), P1000

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.

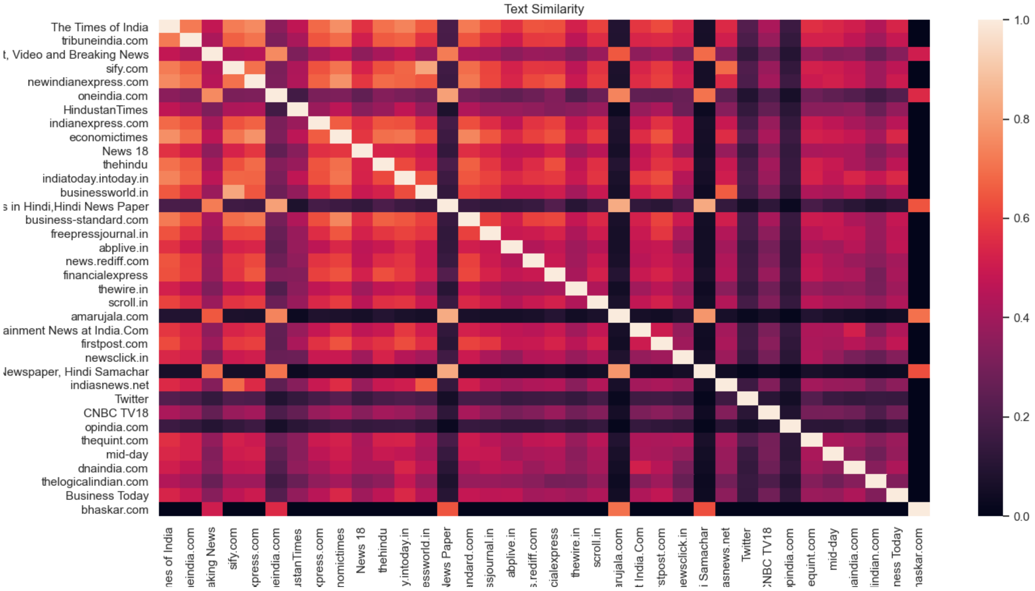
Liu, D., & Chen, Y. (2017). Biterm-LDA: A Recommendation Model for Latent Friends on Weibo. Journal of Residuals Science & Technology, 14(3).

Appendix 1

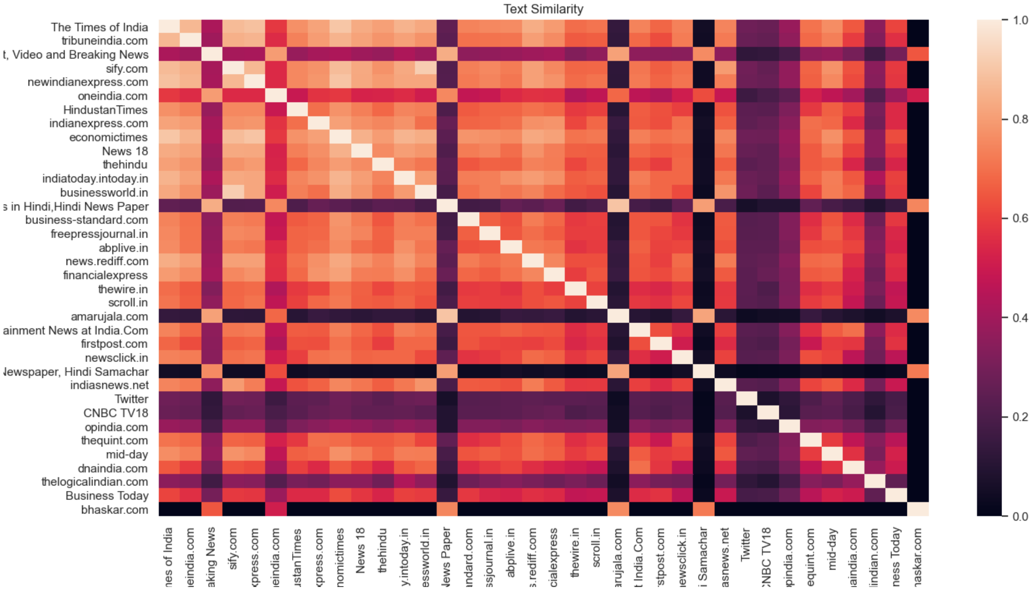
TF-IDF score correlation matrix comparing the text corpus of each group series as whole to highlight the differences how news sources covered the topic.



A1.1 - TF-IDF ranking correlation matrix for Group 1

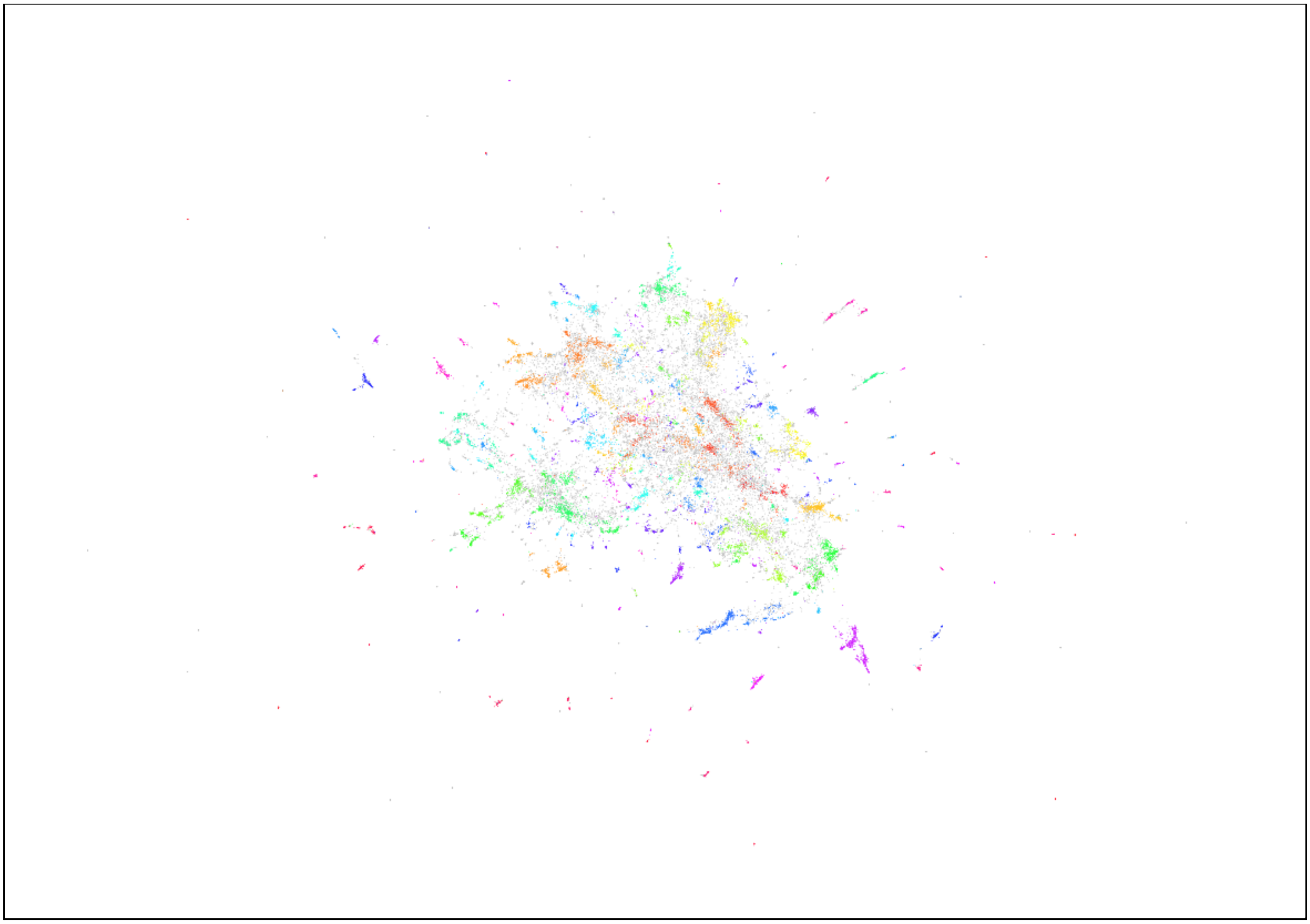


A1.2 - TF-IDF ranking correlation matrix for Group 2



A1.3 - TF-IDF ranking correlation matrix for Group 3

Appendix 2



A2 - All the groups cluster after mBERT, UMAP and HDBSCAN.