2022-06-01

An unsupervised approach to discover media frames

Proceedings of the First Workshop on Natural Language Processing for Political Sciences (PoliticalNLP), pages 22–31
Marseille, 24 June 2022

An Unsupervised Approach to Discover Media Frames

Sha Lai[1], **Yanru Jiang**[4], **Lei Guo**[3], **Margrit Betke**[1], **Prakash Ishwar**[2], **Derry T. Wijaya**[1]
[1] Department of Computer Science, [2] Department of Electrical and Computer Engineering, and
[3] College of Communication, Boston University
[4] Department of Communication, University of California Los Angeles
lais823@bu.edu, yanrujiang@g.ucla.edu, guolei@bu.edu, betke@bu.edu, pi@bu.edu, wijaya@bu.edu

Abstract

Media framing refers to highlighting certain aspect of an issue in the news to promote a particular interpretation to the audience. Supervised learning has often been used to recognize frames in news articles, requiring a known pool of frames for a particular issue, which must be identified by communication researchers through thorough manual content analysis. In this work, we devise an unsupervised learning approach to discover the frames in news articles automatically. Given a set of news articles for a given issue, e.g., gun violence, our method first extracts frame elements from these articles using related Wikipedia articles and the Wikipedia category system. It then uses a community detection approach to identify frames from these frame elements. We discuss the effectiveness of our approach by comparing the frames it generates in an unsupervised manner to the domain-expert-derived frames for the issue of gun violence, for which a supervised learning model for frame recognition exists.

Keywords: unsupervised learning, natural language processing, frame

1. Introduction

Framing, in the communication context, means selecting certain aspect of a perceived reality to improve its salience among the audience (Entman, 1993). By carefully selecting frames, some authors can encourage certain interpretations of an issue; others may even promote a political agenda. Communication researchers have been studying ways to recognize frames in news articles. Currently, their approach is mostly based on manual content analysis. Given an issue, researchers need to create a list of possible frames based on the existing literature and/or by examining a sample of news items. Then human coders are recruited and trained to annotate frames. Such approach has a few limitations. First, since the frames are created about a certain issue, they are limited within some scope. Second, the resulting frames may be subjective, as there is no standard of creating or naming frames. Third, the processes of manually determining and annotating frames can be very time consuming. Though some automatic methods exist have been applied in communication research such as supervised machine learning, they still require substantial expert intervention and human labor. In this article, we propose a framing analysis method that can be widely adapted to different issues, with little human intervention, and largely unsupervised.

Our proposed method, which is our main contribution of this paper, is based on two concepts: general news frames and frame elements. We define general news frames as frames applied to news articles of any issue. To be clear, our proposed "general news frames" are different from "generic news frames" defined in the communication literature such as conflict, economic consequences, and morality (Semetko and Valkenburg, 2000), which are pre-determined by communication scholars. General news frames, as will be discussed later, are identified in an unsupervised way. We define frame elements as ingredients of the general news frames. In communication research, common framing elements include "themes, subthemes, types of actors, actions and setting, qualification, statistics, charts, graphs, appeals, etc" (Van Gorp and others, 2010). Using Wikipedia categories, framing elements in our approach are also identified automatically rather than predetermined. In addition, general news frames are mutually exclusive subsets of all frame elements.

With these two concepts defined, we can describe our proposed approach as a pipeline:

1. Pass a news article to a frame element generator, which makes use of the Wikipedia category system, to obtain a list of frame elements.

2. Pass the list of frame elements to a frame generator, in which we apply graph community detection algorithms, to obtain a list of frames.

A diagram of this pipeline is shown in Figure 1.

2. Related Work

While we study framing in communication research, there exist similar concepts in other domains. In linguistics, for example, semantic frames are defined as a coherent structure of concepts that are related such that without knowledge of all of them, one does not have complete knowledge of any one. Tools like FrameNet (Ruppenhofer et al., 2016) have been developed to recognize these semantic frames from text, but we cannot use them since our task is different due to the difference in definition of frames.

Since our proposed method involves formulation of media frames and the usage of Wikipedia category system, in this section, we will review works related to

computational methods used in media framing research and the application of Wikipedia categories in computational linguistic research.

2.1. Media Framing Analysis

We can categorize the computational methods as the following: lexicon-based methods and machine learning (ML) methods. Furthermore, the ML methods can be split into supervised methods and unsupervised ones. In frame extraction research problems, the frames can either be defined by researchers manually or be modeled and constructed automatically. In the tasks where frames are predefined, a common goal is to recognize the frames from media sources using some lexicon-based or supervised ML methods, while in the event where frames are not explicitly defined, unsupervised ML methods are applied to model them.

2.1.1. Lexicon-based Methods

The lexicon-based methods center around term frequency as well as mapping from keywords to categories. Such methods is widely used in sentiment analysis. For example, Turney (2002) computes similarity between phrases and two lists of predefined words corresponding to positive and negative semantics orientations using Pointwise Mutual Information and Information Retrieval. An example of lexicon-based methods is the development of keywords for frames regarding immigrants by Lind et al. (2019). One major disadvantage of such methods is the requirement of expert knowledge in creating the keywords, which are largely tied to issues, limits the application scope.

2.1.2. Supervised ML Methods

The supervised methods do not gain much popularity in framing analysis, despite of the fact that models like Support Vector Machine (SVM) and Random Forest have been proved successful in other communication research problems (Opperhuizen et al., 2019; Adamu et al., 2021). Burscher et al. (2014) discover that an ensemble algorithm combining two linear SVM models, a polynomial SVM model and a perceptron model can lead to higher accuracy in predicting the four generic frames than using these individual classifiers alone.

Another supervised ML example is the recent work by Tourni et al. (2021), which shows that combining a transformer model to process news headlines and a residual network model to process news images in tandem leads to accurate headline frame prediction.

2.1.3. Unsupervised ML Methods

Despite of the popularity of framing theory in communication research, what constitutes framing remains an open question. Nonetheless, that it being open-ended allows a diverse range of formulation of frames under unsupervised ML approaches.

One popular unsupervised ML method is the Latent Dirichlet Allocation (LDA) based topic modeling. Blei et al. (2003) develop LDA as a probabilistic model that discovers keywords to represent topics in an article. Walter and Ophir (2019) construct frames based on the topics returned by LDA. We would like to emphasize that topics are not equivalent to frames, though they appear to be similar in some cases. One key difference is that frames should be "persistent over time" (Reese et al., 2001) while topics naturally do not have to be so. While we focus on frames in news articles and model them as a collection of frame elements obtained from Wikipedia categories, others may formulate frames in a diverse range of applications. For instance, Ajjour et al. (2019) model frames as mutually exclusive clusters of arguments. They develop a two-level clustering method which takes a set of arguments as input and yields a partition of the arguments as output. Of the two levels of clustering, one aims to remove topics in the arguments and the other aims to produce a partition. Though like us, they model frames as sets, their formulation applies to arguments, which typically contain only one or two sentences and strongly focus on one aspect of the corresponding topic.

2.2. Framing via Community Detection

We construct frames by applying community detection algorithms on a graph formed by frame elements. The community detection has been used extensively in graph analysis and applications. In social science, this technique is frequently applied on social media networks, as a number of reviews and surveys on this type of application have been published (Wang et al., 2015a; Wang et al., 2015b; Bedi and Sharma, 2016; Kumar et al., 2018; Souravlas et al., 2021).

In framing analysis, Walter and Ophir (2019) treat the topics returned by LDA as frame elements. They create a graph using the frame elements and applied community detection on the graph. Such approach is very similar to ours, while the key difference lies in the source of frame elements.

2.3. Usage of Wikipedia Category System

Our approach involves the Wikipedia category system. Many works have adopted this system, but few aim at solving a similar problem as ours. Nastase and Strube (2008) use the system to study the relation between concepts stored on Wikipedia. Pasca (2018) developes a method to recognize classes of Wikipedia articles, where the categories are used as part of the approach. Allahyari and Kochut (2016) integrate the Wikipedia categories as topics into the LDA probabilistic model to perform semantic tagging on online articles.

A number of works use this system to perform topic modeling. Schönhofen (2009) uses Wikipedia categories and Wikipedia article titles to identify document topics. Mirylenka and Passerini (2013) propose a method to create topic summaries for documents by mapping them to Wikipedia articles and the related categories. Kumar et al. (2017) build an automated topic identification model, which is trained on the Wikipedia category graph, to generate topic trees from text data.

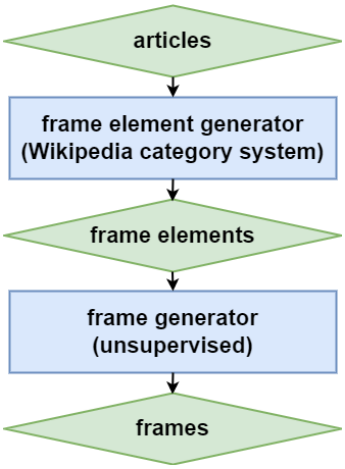

Figure 1: The pipeline of our approach.

Nevertheless, we again stress that topics and frames are different in terms of scope.

3. Methodology

Out proposed pipeline, as described in section 1 and shown in Figure 1, contains two major parts: frame element generator and frame generator. This section presents the two generators in detail.

3.1. Frame Element Generator

The goal of this part is to extract frame elements from the articles. To do so, we first associate each news article to some Wikipedia articles, and then use the Wikipedia category system to create frame elements.

3.1.1. From News Articles to Wikipedia Articles

The bridge between the news articles and the Wikipedia ones is built with computational linguistics techniques. We use a Doc2Vec model (Le and Mikolov, 2014) to create a document embedding for each news article and each Wikipedia one. Then, for each news article embedding, we find the top K_p most similar Wikipedia article embeddings based on cosine similarity. Thus, each news article is linked to K_p Wikipedia ones.

3.1.2. From Wikipedia Articles to Categories

This step involves the category system on Wikipedia. Due to the system's complex nature, we will briefly introduce it with an example before describing how we make use of it.

Wikipedia's Category System Wikipedia is a gigantic online database with free access. For every recorded item, there is a page containing an article describing it. To help the readers better navigate through the database to find relevant items, a hierarchical category system is used to group the articles. Each article has a list of categories, which can be found at the bottom of the article webpage. Furthermore, each category may have its own list of categories.

Example An example that starts from the article "computer science" is illustrated in Figure 2. In this example, the Wikipedia article "computer science" has

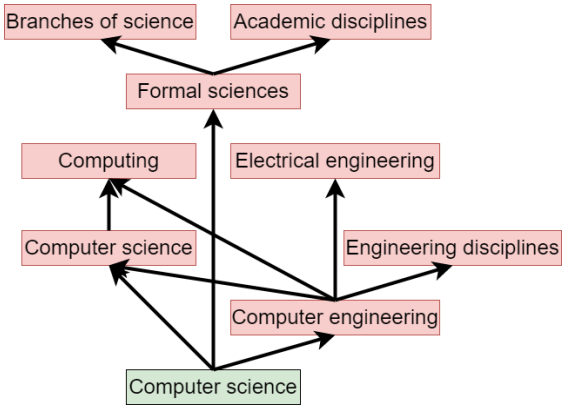

Figure 2: Example Wikipedia categories obtained from two levels of recursion from the Wikipedia article titled "Computer science" (green box).

the following categories: "computer science", "formal sciences", and "computer engineering". The category with the same name, "computer science", has the following categories: "formal sciences", "computing", "categories requiring diffusion", and "commons category link is on Wikidata". The last two of these categories are called "hidden categories" which are mainly used by the Wikipedia's internal system for maintenance purposes. Furthermore, that "formal sciences" is the category of both the article page as well as the category page of "computer science" shows the non-trivial nature of the hierarchical system.

Obtaining Categories The example suggests that one can follow the category links to retrieve the categories recursively starting from any page. Our method performs a recursive retrieval of categories for each page with a maximum recursion depth D.

Processing Categories The retrieved categories need to be cleaned up to reduce noise for further analysis. We first remove the categories for Wikipedia administration or maintenance, including the hidden categories mentioned in the example, since they are not helpful in our study. Then, we merge the categories sharing the same key words via an NLP technique called dependency parsing (DP). DP analyzes the relation between words in a sentence and assigns a grammar role for each word. The main subject is selected uniquely and is called ***root***. All categories are mapped to some ***roots*** and merging occurs among the categories sharing the same ***root***, as these categories document the same subject from different aspects. For example, there are categories such as "suicides by city," "suicides by country," and "suicides by method." In each of them, DP recognizes the word "suicide" as the main subject and labels it ***root***. Then, all three categories can be merged into "suicide."

Forming Frame Elements After processing, we sort the roots by the number of Wikipedia articles they are associated with and choose a list of most popular ones to become frame elements.

3.2. Frame Generator

The frame generator takes the frame elements obtained from the previous steps as inputs and yields frames as outputs. In particular, we build a graph using frame elements and apply graph community detection to partition the frame elements. As a result, each partition will be a frame. In this section, we first define our frame element graph, and then introduce the algorithms used to group the frame elements.

3.2.1. Frame Element Graph

We define a weighted undirected complete graph $G = (V, E, W)$ where the nodes are the frame elements and the edges represent the similarity between the frame elements. The similarity is a combination of two measurement score and is encoded in the edge weight

$$w(u, v) = \textit{ExSim}(u, v) + \textit{SemSim}(u, v).$$

The functions *ExSim* and *SemSim* will be explained next.

ExSim By construction, a frame element is a ***root*** of some Wikipedia categories and the categories are associated with Wikipedia articles. Hence, with an arbitrary ordering of the Wikipedia articles fixed, for each frame element we can define an indicator vector e where $e_i = 1$ if the frame element is associated with the ith article. We call such vector the *existence vector*. The function *ExSim*, where "*Ex*" stands for "existence" and "*Sim*" stands for "similarity", measures the coexistence between two frame elements u and v by computing the cosine similarity of their *existence vectors*.

SemSim Since the frame elements are in the form of text, a natural way to measure their connection is by their semantics meanings. Hence, the function *SemSim*, where "*Sem*" stands for "semantics" and "*Sim*" again stands for "similarity", is added to the weight function. This function computes the cosine similarity between two frame elements' semantics embeddings. To create embeddings, we input the text of each frame element into a pretrained BERT (Devlin et al., 2018) model and extract the outputs of the last layer of the network.

3.2.2. Community Detection

Several algorithms have been developed for different types of graphs, as one algorithm simply cannot perform in all graphs (Javed et al., 2018). We apply two community detection methods : Spectral Clustering (SC), a traditional algorithm that utilizes the eigenvalues and eigenvectors of the graph Laplacian, and Community Discovery via Node Embedding (VEC), a novel method proposed by Ding et al. (2017).

3.3. Summary

We here briefly summarize the relations between the main concepts mentioned so far. Wikipedia categories are reduced and merged into ***roots***, some of which are our frame elements. We build a graph using these frame elements and group them via community detection, and the resulting communities are defined as frames.

Index	Frame
1	2nd Amendment (Gun Rights)
2	Gun Control
3	Politics
4	Mental Health
5	School/Public Space Safety (Public Safety)
6	Race/Ethnicity
7	Public Opinion
8	Society/Culture
9	Economic Consequence

Table 1: The nine headline frames in the Gun Violence Frame Corpus dataset that we used.

4. Experiments

In this section, we will describe the data, our pipeline implementation, the experiments and some intermediate outputs. Since each part of our pipeline involves a number of variables to explore and yields individual outputs, after the data subsection below, we will follow the workflow of the pipeline as in Figure 1 by dividing the subsections similar to the method section. In addition, the code and results are publicly available [1].

4.1. Data

The dataset we used is a subset of the extended Gun Violence Frame Corpus (Liu et al., 2019). The dataset contains 1,300 samples of news articles about gun violence in United States. Each sample has a headline and the main body content. Furthermore, each headline is labeled with one of the nine frames shown in Table 1.

4.2. Frame Element Generator

There are two main steps in the frame element generator we will detail them one at a time below.

4.2.1. From News Articles to Wikipedia Articles

When creating embeddings for our news articles and the Wikipedia ones, we ran a Gensim (Řehůřek and Sojka, 2010) Doc2Vec model and each embedding vector is of length 200, an arbitrary number. There are two common variants of Doc2Vec model: one uses the Distributed Bag of Words version of Paragraph Vector (PV-DBOW) and the other uses the Distributed Memory version of Paragraph Vector (PV-DM) (Le and Mikolov, 2014). PV-DM usually yields more accurate performance in classification tasks while PV-DBOW is faster if the corpus is large. Since the corpus we used to train our Doc2Vec model was a snapshot of all Wikipedia articles taken in June 2021, we chose PV-DBOW for the sake of speed.

For every news article embedding, we found $K_p = 10$ most similar Wikipedia article ones by cosine similarity. A natural way to decide the value to K_p is to examine the overall sorted similarity scores and choose a point where a significant drop locates. However, we observed that the curve went down smoothly from the

[1] https://github.com/slai7880/unsupervised-media-frame-discovery

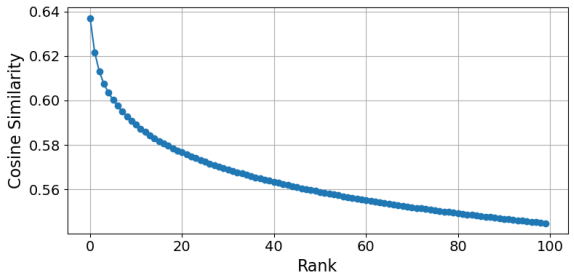

Figure 3: The average top 100 scores of cosine similarity between all news articles and Wikipedia articles. To generate this plot, for each news article, we selected the Wikipedia articles with the 100 highest cosine similarity scores, and, for rank 1 to 100, we averaged the scores across articles. The average scores do not have a sudden decrease in this range.

highest to the 100th, as depicted in Figure 3. Thus, an arbitrary number 10 was chosen for this part.

4.2.2. From Wikipedia Articles to Categories

This part involves retrieving and cleaning categories.

Obtaining Categories As described in section 3.1.2, we retrieved categories recursively with a maximum depth D. In our implementation, we set $D = 4$. Our observation of the Wikepedia category system suggested that if D is too small, the retrieved categories might be too specific, while our communication experts recommended more general categories for better framing quality. Furthermore, as shown in Figure 2, since each category can also have a list of its own categories, the farther the exploration goes, the more categories to examine. This means the time it takes to retrieve the categories can increase drastically. Therefore, in this study, we fixed this upper bound D to be 4, with which the retrieval process could finish in a reasonable amount of time and the outcomes were deemed satisfactory by our communication experts.

Processing Categories The category retrieval process returned 74,281 categories. After the administrative and maintenance categories were removed, 71,303 remained. We then applied a dependency parser developed by Qi et al. (2020) to obtain 4,797 root words. Next, we sorted the root words by the size of the union of the directly associated Wikipedia articles and chose the top 100 root words to become our frame elements. We consider a root word and a Wikipedia article is directly associated if the root word is a category of the article or one of the categories of the article is merged into the root word by dependency parsing.

4.3. Community Detection

With the selected root words being our frame elements, we began building the graph as described in 3.2.1 for community detection. In this phase, we explored different numbers of communities, as there is no common method to predetermine the right value for this parameter. More specifically, for each integer N_c between 2

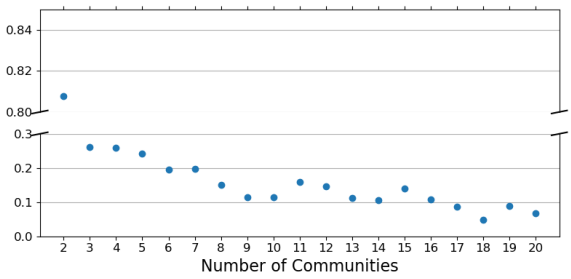

Figure 4: Adjusted Rand index score between the community labels produced by SC and VEC.

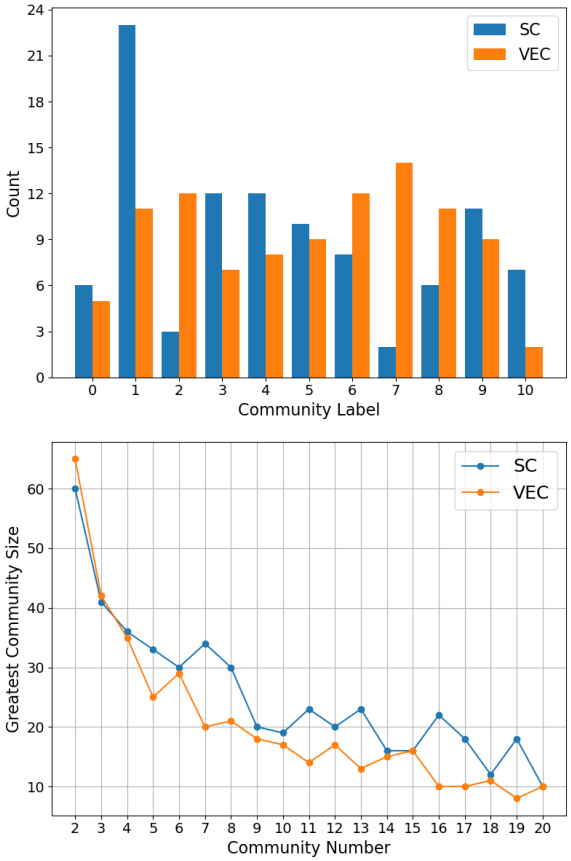

Figure 5: Top: The number of frame elements in each community with total community number N_c fixed to be 11. For this example, SC produces a dominant communit. Bottom: The greatest community sizes for each community number from 2 to 20. SC tends to produce a dominant community while VEC is less likely to do so, as the VEC curve is mostly below the SC one.

and 20, we ran the two community detection algorithms with the community number set to N_c.
We examined the community detection results both from data science and communication perspectives.

4.3.1. Analysis from Data Science Perspective

Our first impression is that the clustering results are very different between the two algorithms for any community number. This can be verified by adjusted Rand index (Hubert and Arabie, 1985) shown in Figure 4, where the majority of the values appear to be very low. Furthermore, we observe that SC tends to produce a

community with size dominating the results. For example, as seen in the top plot in Figure 5, among the communities produced by SC, community 1 dwarfs the rest by size. Such situation, however, is not seen in the results produced by VEC. The bottom plot in Figure 5, where we show the maximum community size in each community number setting, suggests that such dominating community is common in SC results.

4.3.2. Analysis from Communication Perspective

Our communication experts examined the results by determining how coherent and how interpretable each community is. In particular, a group of frame elements are coherent if they are distinct from each other and semantically meaningful, and elements within a cluster represent a core frame (Guo et al., 2016; Van Gorp and others, 2010).

Overall, a good range for the community number appears to be between 7 and 16 for both SC and VEC. The results with community number $N_c < 7$ are too broad to identify meaningful frame clusters, while the ones with $N_c > 16$ are too sporadic.

SC tends to outperform VEC for all community numbers, particularly in terms of coherency, despite of the presence of the dominating community. In fact, most communities from SC are coherent and interpretable except for the dominating ones. Interestingly, since usually smaller communities are more coherent than larger ones, the existence of dominating communities, which results in the smaller ones in the same set of outputs, is likely the reason why SC is overall better than VEC. Furthermore, the best community numbers, judging from SC results, are 12, 14, 15, and 16, with 12 and 14 being slightly better.

5. Evaluation

The final part is to evaluate the community frames. However, the evaluation is not simple, because the communities do not have labels and neither do the news contents. Nevertheless, we devised an evaluation approach that made use of the nine headline frames.

Our evaluation strategy aims to create "soft labels" based on the nine headline frames for both the articles and the communities. Thus, this requires us to first obtain frames from the main body of each article and then associate the communities to the nine frames. The first part was achieved by predicting the frame for each sentence in the articles while the second part was achieved by associating the communities to the nine headline frames. We applied a BERT model in both parts, but in each part the model and usage were different. We will detail them one at a time below.

5.1. Acquiring Article Frames

Since it has been shown by Tourni et al. (2021) that BERT can accurately predict the frames of the headlines in the same dataset we used, we adopted the model and a training process similar as in that work.

Training BERT We created a training set for BERT using 2,911 news headlines from the Gun Violence Frame Corpus dataset. Among these headlines, 1,300 were the samples we used in our current framing analysis and each of them has one of the nine frames as listed in Table 1, while the rest were labeled by our communication experts as "no frame". We assumed that the set of frames present in the articles are the same as those in the headlines, and that there exist a substantial number of sentences not having one of the nine frames. In fact, many sentences do not even have a frame. An example is a quote from a conversation. We finetuned the epoch number and the learning rate using stratified 5-fold cross validation. The optimal values for these two parameters are 12 and 2×10^{-5} respectively, and the corresponding optimal validation F1 score is 0.817.

Preprocessing Articles Before predicting the sentence frames, we first tokenized each article into sentences using the NLTK (Loper and Bird, 2002) package, and then we removed sentences of length less than 20 characters, since we observed that most sentences shorter than 20 characters did not contain any frame. After removal, we found that every news article had at least one sentence left while the majority (77%) still had more than 10 sentences left.

Prediction Outputs The prediction results are shown in Figure 6. An interesting observation from these histograms is that the distribution shape of the frames in the prediction roughly resembles that in the training set.

Creating Soft Labels Finally, we created a soft label L_s for each article where the ith entry $L_s(i)$ is the proportion of sentences in the article that were classified as headline frame i. These soft labels would serve as ground truth.

5.2. Community-Frame Association

In this part, we need to associate the communities to the nine headline frames. We again use BERT.

Creating Community Embeddings For each frame element, we built a BERT embedding by extracting the outputs of the last layer of the model. Note that the BERT used for this task was pretrained but not finetuned on any problem. Then, we computed the embedding centroid by averaging all frame element embeddings in each community. Next, for each centroid i, we computed a vector S_i where each entry s_{ij} is the cosine similarity between centroid i and headline frame embedding j. This gave us a measurement of how close each community is to the nine known frames.

Creating Soft Labels Because every news article is linked to a list of communities, for each news article, we computed the average of the similarity vectors corresponding to the communities linked to the article and then normalized the resulting vector. The final output vector, denoted as L_c, would be the soft label of the article from community detection.

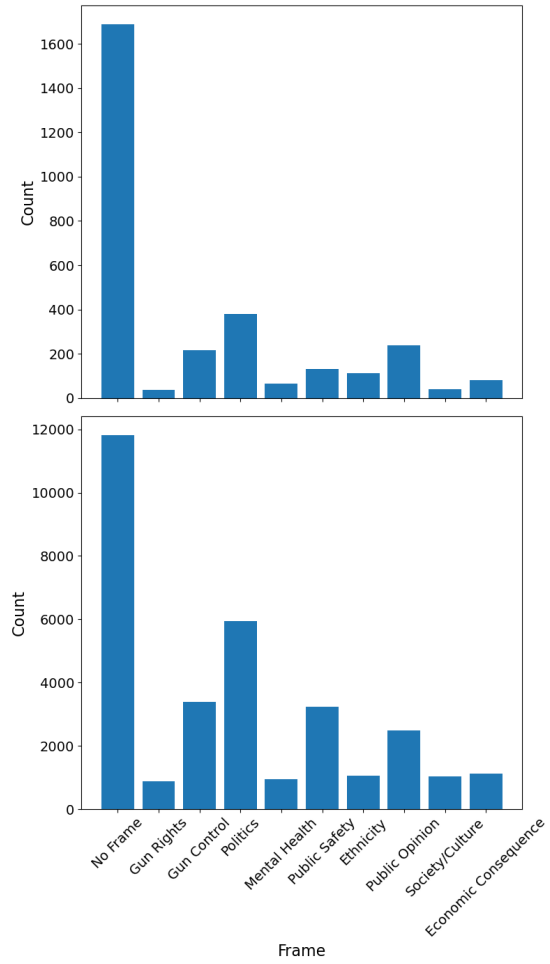

Figure 6: Top: Frame population among the headlines used to train BERT. Bottom: The sentence frame distribution predicted by BERT.

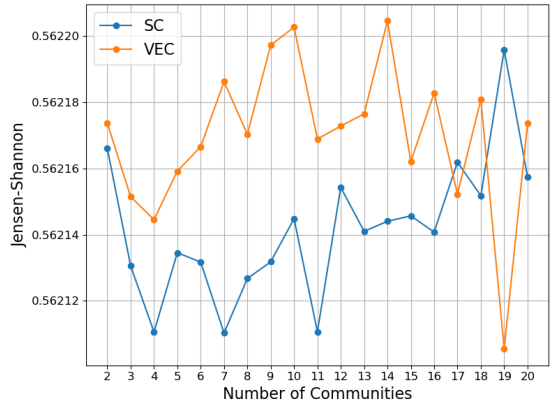

Figure 7: The average Jensen-Shannon distance.

5.3. Comparing Soft Labels

The evaluation is to compare the soft labels from the two sources described above. We present in Figure 7 the results in Jensen-Shannon distance. In the case of SC, there are three values for the number of communities N_c where the distance is minimized: 4, 7 and 11. Whereas there is only one obvious minimum for VEC: $N_c = 19$. It's interesting that the minimum of the VEC curve is exactly the maximum of the SC one.

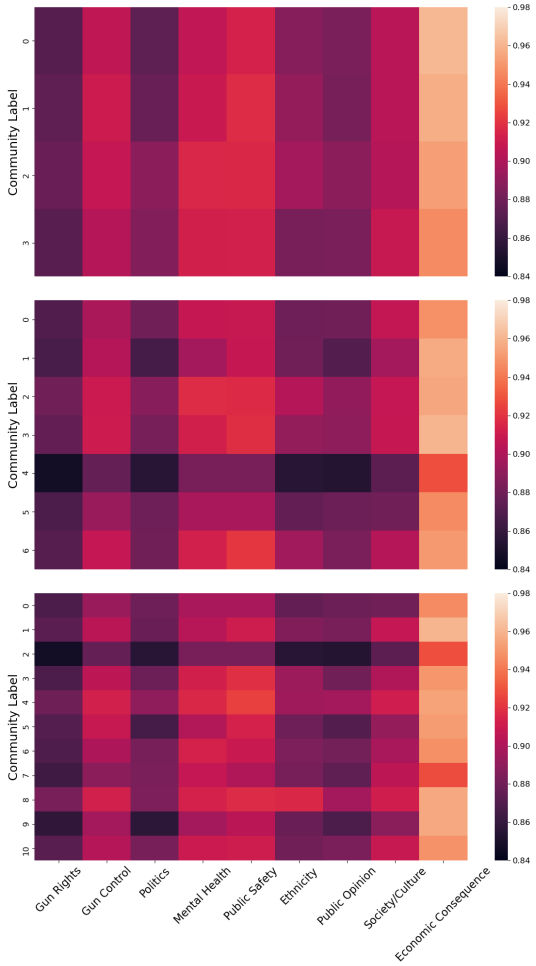

Figure 8: From top to bottom: soft labels corresponding to $N_c = 4$, $N_c = 7$ and $N_c = 11$. Note that the community labels are zero-based.

5.4. Examining Soft Labels

Because each soft label of a community is a vector of similarity towards the headline frames, we can visualize the soft labels using heatmap as shown in Figure 8 and in Figure 9. In particular, Figure 8 shows the soft label heatmaps corresponding to $N_c = 4, 7$ and 11 on the SC curve, and Figure 9 shows the soft label heatmap at $N_c = 19$ on the VEC one.

A common feature among these figures is that the frame ***Economic Consequence*** always has the highest similarity score towards any community, regardless of the community detection method used. In fact, we observe such dominance in the rest of the results as well. However, as shown in Figure 6, according to BERT, the frame ***Economic Consequence*** is among the low-popularity frames. A similar and more surprising phenomenon can be observed in frames ***Politics*** and ***Public Opinion***, which are both popular in the predicted sentence frames but almost always have low similarity scores towards the communities.

After examining the frame elements and their original Wikipedia categories, we found a possible cause of such difference for frame ***Politics***: many words that are apparently related to this frame were dropped by

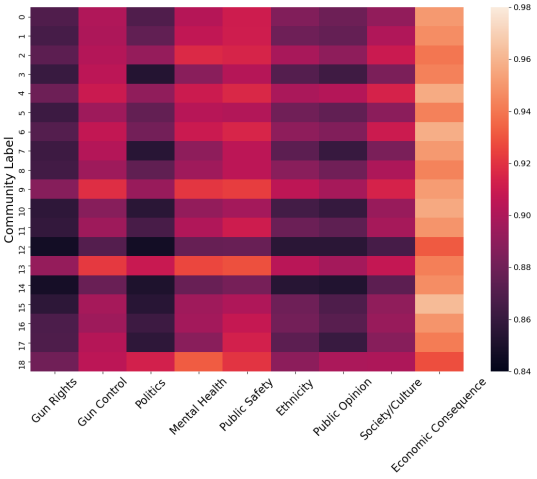

Figure 9: The soft label heatmap of the VEC results where $N_c = 19$.

DP. For example, the category "Political positions of United States senators", which is obviously related to ***Politics***, was reduced by the parser to "position", which has little similarity to that frame. Such loss of information appears to be a limitation of applying DP.

6. Discussion

The pipeline we propose can be beneficial in both the communication and computational linguistic fields.

In communication research, lexicon-based methods and supervised ML are most commonly used in automatic framing analysis. As mentioned in 2.1.1, the nature of lexicon-based methods requires researchers to create keywords and dictionaries to map the keywords to frames. Supervised ML also requires human annotations. Our pipeline, however, requires much less labor, as the process is unsupervised. Furthermore, since the frame elements we choose are essentially Wikipedia categories, they are not tied to any specific issues. Hence, the frames constructed using these frame elements are more general and can be applied to articles of any issue. We, however, recommend that researchers should consider using our method as an exploratory approach examine the text rather than use it for hypothesis testing. More future research should be conducted to test the validity of the proposed approach.

In computational linguistic research, our idea of forming frame elements based on Wikipedia categories adds another novel usage of this gigantic knowledge database. Other researchers can use a similar approach to formulate abstract concepts from text like we do with Wikipedia categories. Our proposed evaluation method can be an example of using a pretrained model to create ground truth information for comparison when such information does not exist in some scenario. In addition, the performance presented in Figure 7 can serve as a baseline for future unsupervised automatic framing research. Furthermore, since Wikipedia is a multilingual knowledge database, we can adopt our pipeline in analyzing text in non-English languages.

Our work can also be applied on text other than news articles. For instance, as pointed out by Odebiyi and Sunal (2020), some textbooks used in U.S. schools seem to be portraying Africa nations falsely. The authors approach this framing problem in textbooks by analyzing themes. More specifically, they identify three main categories of themes: 1) the framing of Nigeria(ns), which includes a) "resources and poverty" and b) underdevelopment and conflicts as sub-frames; 2) demographic features and framing of Nigeria(ns); and 3) cultural practices (mis)understanding and ecological framing of Nigeria(ns). The process is done through three rounds of manual coding. First round, the coders locate the text relevant to Nigeria. Second, the coders examine the relevant portions sentence by sentence. Third, the coders "conduct focused coding to create meta-codes for different patterns and themes based on how each textbook framed Nigerian people, places and practices" found in the previous rounds. Such process is time consuming and heavily human-labor involving. If we adopt our proposed pipeline into this problem, we can simplify the work by inputting the textbook articles into our model, as we do with news articles, and obtaining the frames as clusters of frame elements. Some post-processing work may be required, as the clusters by themselves do not have names. Moreover, the thematic approach used by the authors is bound to the specific nation, while ours can produce more robust frames that can extend the analysis to more Africa countries.

7. Conclusion and Future Work

In this work, we have presented a novel unsupervised pipeline method to produce frames for news articles. We have proposed using Wikipedia categories to create frame elements. We have formulated the frame construction as a graph community detection problem where the frame elements serve as graph nodes. We have demonstrated an example of our pipeline using the news from Gun Violence Frame Corpus. Lastly, we have proposed an evaluation strategy to compare our community frames and the news article ones.

Automatic framing, especially when pairing with an unsupervised method, remains a challenging task. Our future work involves improvement and exploration in many steps of our pipeline. In particular, we seek better handling of the Wikipedia categories, as simply merging them by dependency parsing can result in loss of helpful information. Furthermore, since the two community detection methods we used only assign one community label for each graph node, we plan to explore methods that allow multiple labels.

Acknowledgements Funding from the National Science Foundation, grant 1838193, is gratefully acknowledged.

8. Bibliographical References

Adamu, H., Lutfi, S. L., Malim, N. H. A. H., Hassan, R., Di Vaio, A., and Mohamed, A. S. A. (2021). Framing twitter public sentiment on nigerian government covid-19 palliatives distribution using machine learning. Sustainability, 13(6):3497.

Ajjour, Y., Alshomary, M., Wachsmuth, H., and Stein, B. (2019). Modeling frames in argumentation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP/IJCNLP), pages 2922–2932. pdf.

Allahyari, M. and Kochut, K. (2016). Semantic tagging using topic models exploiting wikipedia category network. In 2016 IEEE Tenth International Conference on Semantic Computing (ICSC), pages 63–70.

Bedi, P. and Sharma, C. (2016). Community detection in social networks. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 6(3):115–135.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan):993–1022.

Burscher, B., Odijk, D., Vliegenthart, R., De Rijke, M., and 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.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805.

Ding, W., Lin, C., and Ishwar, P. (2017). Node embedding via word embedding for network community discovery. IEEE Transactions on Signal and Information Processing over Networks, 3(3):539–552.

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. Journal of communication, 43(4):51–58.

Guo, L., Vargo, C. J., Pan, Z., Ding, W., and Ishwar, P. (2016). Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. Journalism & Mass Communication Quarterly, 93(2):332–359.

Hubert, L. and Arabie, P. (1985). Comparing partitions. Journal of classification, 2(1):193–218.

Javed, M. A., Younis, M. S., Latif, S., Qadir, J., and Baig, A. (2018). Community detection in networks: A multidisciplinary review. Journal of Network and Computer Applications, 108:87–111.

Kumar, S., Rengarajan, P., and Annie, A. X. (2017). Wikitop: Using wikipedia category network to generate topic trees. In Thirty-First AAAI Conference on Artificial Intelligence.

Kumar, P., Chawla, P., and Rana, A. (2018). A review on community detection algorithms in social networks. In 2018 4th International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), pages 304–309. IEEE.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In International conference on machine learning, pages 1188–1196. PMLR.

Lind, F., Eberl, J.-M., Heidenreich, T., and Boomgaarden, H. G. (2019). Computational communication science— when the journey is as important as the goal: A roadmap to multilingual dictionary construction. International Journal of Communication, 13:21.

Liu, S., Guo, L., Mays, K., Betke, M., and Wijaya, D. T. (2019). Detecting frames in news headlines and its application to analyzing news framing trends surrounding U.S. gun violence. In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), pages 504–514, Hong Kong, China, November. Association for Computational Linguistics.

Loper, E. and Bird, S. (2002). Nltk: The natural language toolkit. CoRR, cs.CL/0205028.

Mirylenka, D. and Passerini, A. (2013). Navigating the topical structure of academic search results via the wikipedia category network. In Proceedings of the 22nd ACM international conference on Information & Knowledge Management, pages 891–896.

Nastase, V. and Strube, M. (2008). Decoding wikipedia categories for knowledge acquisition. In AAAI, volume 8, pages 1219–1224.

Odebiyi, O. M. and Sunal, C. S. (2020). A global perspective? framing analysis of us textbooks' discussion of nigeria. The Journal of Social Studies Research, 44(2):239–248.

Opperhuizen, A. E., Schouten, K., and 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.

Pasca, M. (2018). Finding needles in an encyclopedic haystack: Detecting classes among wikipedia articles. In Proceedings of the 2018 World Wide Web Conference, pages 1267–1276.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations.

Reese, S. D., Gandy Jr, O. H., and Grant, A. E. (2001).

Framing public life: Perspectives on media and our understanding of the social world. Routledge.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pages 45–50, Valletta, Malta, May. ELRA. http://is.muni.cz/publication/884893/en.

Ruppenhofer, J., Ellsworth, M., Schwarzer-Petruck, M., Johnson, C. R., and Scheffczyk, J. (2016). Framenet ii: Extended theory and practice. Technical report, International Computer Science Institute.

Schönhofen, P. (2009). Identifying document topics using the wikipedia category network. Web Intelligence and Agent Systems: An International Journal, 7(2):195–207.

Semetko, H. A. and Valkenburg, P. M. (2000). Framing european politics: A content analysis of press and television news. Journal of communication, 50(2):93–109.

Souravlas, S., Sifaleras, A., Tsintogianni, M., and Katsavounis, S. (2021). A classification of community detection methods in social networks: a survey. International Journal of General Systems, 50(1):63–91.

Tourni, I., Guo, L., Daryanto, T. H., Zhafransyah, F., Halim, E. E., Jalal, M., Chen, B., Lai, S., Hu, H., Betke, M., Ishwar, P., and Wijaya, D. T. (2021). Detecting frames in news headlines and lead images in U.S. gun violence coverage. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4037–4050, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Turney, P. D. (2002). Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. arXiv preprint cs/0212032.

Van Gorp, B. et al. (2010). Strategies to take subjectivity out of framing analysis. Doing news framing analysis: Empirical and theoretical perspectives, pages 84–109.

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

Wang, C., Tang, W., Sun, B., Fang, J., and Wang, Y. (2015a). Review on community detection algorithms in social networks. In 2015 IEEE international conference on progress in informatics and computing (PIC), pages 551–555. IEEE.

Wang, M., Wang, C., Yu, J. X., and Zhang, J. (2015b). Community detection in social networks: an in-depth benchmarking study with a procedure-oriented framework. Proceedings of the VLDB Endowment, 8(10):998–1009.