Group 04: Explainable Topic Model for Fact-Checked Claims

Aditya Singhal - 1154832 **Anwar As'ad - 1172339** **Bithy Das - 1159022**

Hasib Kamal - 1155165 **Mohammad Hossein Ansari - 1152091**

Abstract

With exponential increase in the amount of content available on web, internet users spend long hours in finding relevant information. This project aims to explore a solution to help users spend less time on the searching process using topic modeling. We implement four different topic models, Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF), Latent Semantic Indexing (LSI), and Hierarchical Dirichlet Processes (HDP) on a public health dataset, and measure their performances using two different coherent scores: *c_v*, and *c_umass*. Then, we create different propagation graphs to visually explain inter-topic correlation between the generated topics. Results indicate that LSI is the most suitable topic modeling method for our use-case, and we discuss some interesting findings.

1 Introduction

Finding relevant information, especially during a pandemic situation is a fibrous task. The problem becomes more complex when users are interested in focusing on content from a particular field. For the scope of this project, we focus our work on conducting topic modeling on fact-checked health claims, which consist of different factual claims made by prominent leaders in recent years.

It is possible to identify the topic and label it by using different techniques. Topic models are useful for a wide range of tasks, such as text classification and trend detection. It is a versatile approach to analyzing unstructured texts. Generally, the learning process is unsupervised, but this project uses a semi-supervised learning approach. By treating the instances as a mixture of topics, we generate a cluster of keywords. Figure 1 shows a brief and simple representation of the input/output structure. First, the model takes a set of claims as input, pre-processes the text, and detects word and phrase patterns within them using different clustering techniques, thereby automatically clustering word groups and similar expressions that characterize a set of claims. Then, the topics from best performing model are used to propose suitable topic names. The claims can have a variety of features which are examined, and the resulting topic is a sequence of words used to generate propagation graphs to make the model explainable. This project contributes to existing work by proposing a method for users to effectively analyze large textual content using existing topic modeling methods, and visualize them using different propagation graphs to explain the inter-topic correlation.

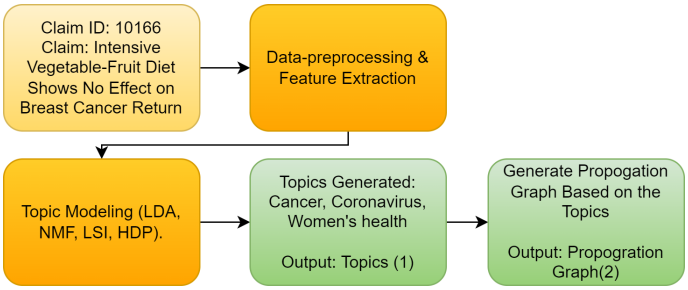

Figure 1: Input/Output Example of the System

The rest of the paper is organized as follows. Section 2 presents background on the previous studies conducted. Section 3 presents an overview of dataset used for this project. Section 4 introduces our approach in detail. Later, in section 5 we present the results. Finally, Section 6 concludes the paper.

2 Related Work

Recent works on explainable fact-checking claims for medical fields have used summarization models to generate an explanation for a text input Kotonya and Toni (2020). While that is a possible solution to the problem, the generated output of summarization models can still sometimes be somewhat long causing potential inconvenience to some users.

In addition, summarization models tend to miss important details and produce unreliable outputs, which can have serious repercussions in the medical domain. Therefore, this work intends to utilize topic modeling instead. With topic modeling, it is possible to produce a variety of topics that cover the general ideas of a corpus and generate visuals using these topics to help users understand the context of a corpus and how its topics correlate with each other at a glance. In other words, the goal is to find the best model for generating topics on this dataset and use the topics to explain the correlation between them via propagation graphs.

Each model and algorithm is fit for a certain condition. To understand which fits our task the best, we implement various models. For instance, LDA is a probabilistic model with interpretable topics. However, it has the disadvantage of generating topics as soft-clusters Onan et al. (2016). NMF generates a semantically meaningful result that is easily interpretable in clustering applications. It has been widely used as a clustering method, especially for document data, and as a topic modeling method Kuang et al. (2015). HDP holds the advantage of generating the maximum number of topics that can be unbounded and learned from the dataset instead of being defined beforehand Bisk and Hockenmaier (2013). LSI lacks the ability to shard and map-reduce, and tends to have lower accuracy compared to LDA. However, it helps overcome synonymy by increasing recall Ramamonjisoa (2014).

3 Dataset

PUBHEALTH [1] Kotonya and Toni (2020) is a comprehensive dataset for explainable automated fact-checking of public health claims. Since the project involves semi-supervised learning for topic modeling, we merge all the training, testing, and validation claims together, resulting in 14,752 short text claims also known as instances.

4 Approach & Methodology

4.1 Topic Modeling

Pre-processing: First, the claims obtained from the dataset are cleaned. All instances are converted to lowercase, followed by removal of all non-alphabets (punctuation, numbers, new-line characters and extra-spaces). Next, we remove all URLs and extra-spacing. Then English stopwords, using online library *(nltk library [2])*, along with Medical stopwords Ganesan et al. (2016) were filtered out to remove low-level information from the claims to give more importance to the remaining words Pradana and Hayaty (2019). Small words (length < 3) were removed as they cause complications and unreliability with model's output. Stemming and lemmatization were also applied onto the claims to reduce inflectional words and convert derivationally related forms of a word to a common base word to help with the model's output. To finish the pre-processing, tokenization was performed on the claims to break down the short texts into small chunks which are called tokens. Tokenization helps models by interpreting the meaning of the tokens by analyzing the sequence of the tokens Mielke et al. (2021).

Feature extraction: Features form the tokenized claims were extracted using Term Frequency – Inverse Document Frequency also known as TF-IDF. It is a statistical measure which works by evaluating the relevance of a word in a document in a collection of documents. It works multiplying how many times a word appears in a sentence (Term Frequency) and the inverse frequency of the word across a collection of documents (Inverse Document Frequency) Kim et al. (2019). Grid search was used to determine the optimal hyper-parameters for training the models Liashchynskyi and Liashchynskyi (2019). With the acquisition of optimal hyper-parameters 4 topic modeling models were trained using the tokenized claims.

Modeling: Next, we pass the features as input to various models from gensim library[3]. Latent Dirichlet Allocation (LDA) is a topic modeling model which labels topics based on word frequency from a set of documents. It is used for discovering hidden grouping within documents and works well in large dataset. It gives multiple topics for each document and more generalized in terms of topic generation Jelodar et al. (2018). Latent Semantic Analysis (LSI) uses Singular Value Decomposition (SVD) a mathematical operation which reduces the input to its core for simple and efficient calculation Lossio-Ventura et al. (2021). Non-Negative Factorization (NMF) reduces the dimension of the input and uses factor analysis method to give comparatively less weight to the words that have less coher-

[1] https://github.com/neemakot/Health-Fact-Checking

[2] https://www.nltk.org/

[3] https://radimrehurek.com/gensim/

ence. It is becoming popular for its ability to extract sparse and interpretable features. It works well with TF-IDF normalized documents O'Callaghan et al. (2015). Hierarchical Dirichlet Process (HDP) is an extension of LDA which was designed to work where the number of topics in a document is not known. HDP uses Dirichlet process to calculate the uncertain number of topics and learns the number of topics form the data Altaweel et al. (2019).

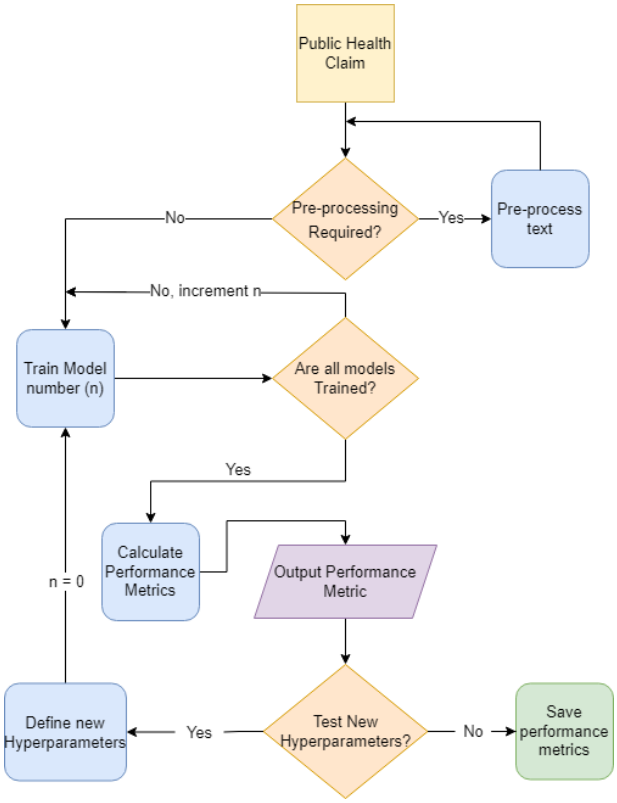

Figure 2: Code Structure

We measure their performances by calculating two coherent scores: *c_v* and *c_umass* scores. *c_v* uses one-set segmentation of top words and an indirect confirmation measure through normalized pointwise mutual information and cosine similarity. Whereas *c_umass* uses one-preceding segmentation and confirmation measure through logarithmic conditional probability Kapadia (2020). Generally, the higher the score, the better a model is considered to perform. For baseline models, previous research on similar tasks shows *c_v* and *c_umss* scores to range between: (0.37 to 0.52) and (-2.5932 to -12.4762) respectively, and we try to maximize both Dahal et al. (2019). We also conduct a grid search to find the best hyper-parameters for these models. This help us understand the coherence of the topic modeling for each implementation. Figure 2 illustrates the flow of our code structure.

4.2 Propagation Graphs

The Networkx [4] library was utilized to create propagation graphs and explain inter-correlation between topics and dataset. We filter the dataset, and keep only the rows that contain one of the keywords belonging to a topic. We store the frequency of occurrence for the topic in dataset as weight of each edge.

To create inter-topic propagation graph, we perform an inner join to find common keywords' occurrence in each instance, and store the the frequency as weight of edge between the two topic nodes.

Algos	run	state	c_v	c_umass	time(s)
LDA	1	2	0.42	-10.44	589.07
NMF	1	2	0.30	-7.22	140.06
NMF	2	4	0.31	-7.46	140.35
LSI	1	2	0.27	-6.54	3.98
LSI	2	4	0.27	**-6.44**	3.96
HDP	1	2	**0.75**	-20.02	37.56
HDP	2	4	0.75	-19.98	37.48

Table 1: Coherence Scores for all Models

Topic Keywords	Topic
['cancer', 'study', 'drug', 'test', 'risk', 'breast', 'say', 'show', 'health']	Cancer
['health', 'coronavirus', 'drug', 'trump', 'state', 'death']	Coronavirus
['test', 'case', 'woman', 'report', 'plan', 'study']	Women's health
['heart', 'breast', 'show', 'find', 'prostate', 'health', 'attack']	Emergency health conditions
['trump', 'president', 'donald', 'mental', 'promise']	US Presidency
['year', 'blood', 'find', 'approve', 'promise']	Blood donation
['health', 'drug', 'test', 'study', 'year', 'show', 'cancer']	Drug tests
['coronavirus', 'trump', 'people', 'vaccine', 'donald', 'risk']	US & Coronavirus
['heart', 'health', 'attack', 'cancer']	Cardiac disease
['help', 'risk', 'treatment', 'show', 'disease', 'patient']	Disease treatment

Table 2: Topics Generated Using LSI Model

[4] https://networkx.org/

5 Results

Table 1 shows the coherence scores for each of the algorithm on the complete dataset. While HDP gives the highest *c_v score*, LSI gives highest *c_umass* scores and generates most meaningful topics (Table 2). The top-10 topic keywords generated are chosen to search for on the first page of Google search results. The resulting contents are then retrieved to interpret the extracted topic keywords to propose a suitable topic name. For example, for the set of keywords yielded by the topic model: *['cancer', 'study', 'drug', 'test', 'risk", 'breast', 'say', 'show', 'health']*, we assign it the topic: *'Cancer'*.

The first propagation graph (Figure 3) is populated with topics as outer nodes and the dataset as the central node (Table 3). The width of the edges indicates the correlation between the topics and the dataset. It is evident that the correlation between the US Presidency and the dataset is the least, while Cancer has the highest correlation.

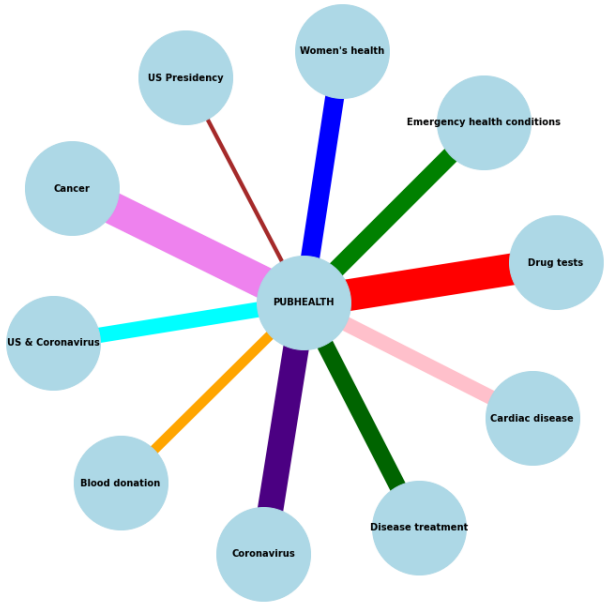

Figure 3: Propagation graph (I) showing the correlation between each topic and the PUBHEALTH dataset

The second propagation shows inter-topic correlation between all topics generated using LSI model. It assesses the level of relation between any two topics in the dataset. Figure 4 shows Cancer and Drug Tests have the highest correlation between them, creating the thickest edge. However, topics like US Presidency and Women's Health have the lowest correlation, which is reflected in weight value 47, and explains the thinnest edge between them. This is also reflected by considering real factors as these two topics are have far less in common with each other than any other topic. We find also strong correlation between Coronavirus and Cardiac disease. Refer Appendix for edge weights of the propagation graphs.

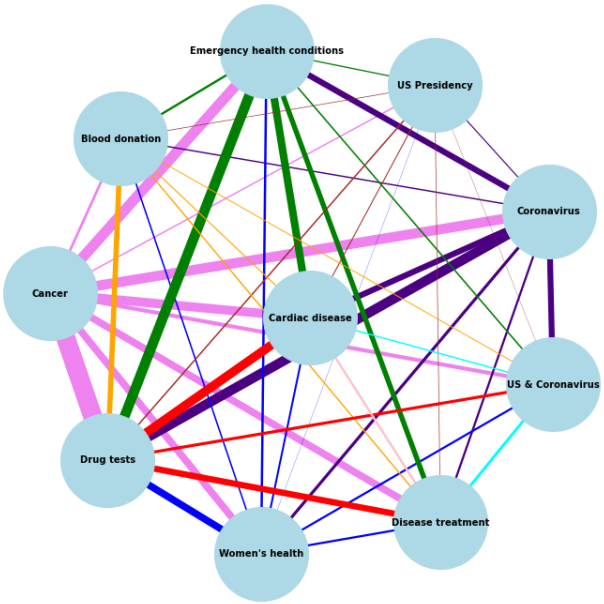

Figure 4: Propagation graph (II) showing the inter-topic correlation

6 Conclusion and Discussion

Topic modeling has great potential in reducing the amount of time users spend when researching information. Considering domain specific medical keywords for stopword removal and converting an unsupervised topic model to a semi-supervised model not only gives us a good idea of the topics discussed in a corpus, but also helps generate visualizations via propagation graphs to understand the correlation of topics discussed at a glance. This model can be applied to generate any kind of label on a document, such as a tag on a website post. The use of word proximity in documents also helps maintain the integrity of the topic.

While this work focuses on topic modeling using four benchmark methods, future work can be expanded to explore more methods in this domain. In addition, our work focuses on health-text domain. It would be interesting to explore how it performs in others. We use propagation graphs to create visualizations and explain the topics generated. Future research can explore other methods for making models explainable.

References

Mark Altaweel, Christopher Bone, and Jesse Abrams. 2019. Documents as data: A content analysis and topic modeling approach for analyzing responses to ecological disturbances. *Ecological informatics*, 51:82–95.

Yonatan Bisk and Julia Hockenmaier. 2013. An hdp model for inducing combinatory categorial grammars. *Transactions of the Association for Computational Linguistics*, 1:75–88.

Biraj Dahal, Sathish A. P Kumar, and Zhenlong Li. 2019. Topic modeling and sentiment analysis of global climate change tweets. *Social network analysis and mining*, 9(1):1–20.

Kavita Ganesan, Shane Lloyd, and Vikren Sarkar. 2016. Discovering related clinical concepts using large amounts of clinical notes: Supplementary issue: Big data analytics for health. *Biomedical engineering and computational biology*, 7s2:BECB.S36155–.

Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. 2018. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia tools and applications*, 78(11):15169–15211.

Shashank Kapadia. 2020. Evaluate topic models: Latent dirichlet allocation (lda).

Donghwa Kim, Deokseong Seo, Suhyoun Cho, and Pilsung Kang. 2019. Multi-co-training for document classification using various document representations: Tf–idf, lda, and doc2vec. *Information sciences*, 477:15–29.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Da Kuang, Jaegul Choo, and Haesun Park. 2015. Nonnegative matrix factorization for interactive topic modeling and document clustering. In *Partitional Clustering Algorithms*, pages 215–243. Springer.

Petro Liashchynskyi and Pavlo Liashchynskyi. 2019. Grid search, random search, genetic algorithm: A big comparison for nas.

Juan Antonio Lossio-Ventura, Sergio Gonzales, Juandiego Morzan, Hugo Alatrista-Salas, Tina Hernandez-Boussard, and Jiang Bian. 2021. Evaluation of clustering and topic modeling methods over health-related tweets and emails. *Artificial intelligence in medicine*, 117:102096–102096.

Sabrina J Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y Lee, Benoît Sagot, and Samson Tan. 2021. Between words and characters: A brief history of open-vocabulary modeling and tokenization in nlp.

Aytug Onan, Serdar Korukoglu, and Hasan Bulut. 2016. Lda-based topic modelling in text sentiment classification: An empirical analysis. *Int. J. Comput. Linguistics Appl.*, 7(1):101–119.

Derek O'Callaghan, Derek Greene, Joe Carthy, and Pádraig Cunningham. 2015. An analysis of the coherence of descriptors in topic modeling. *Expert systems with applications*, 42(13):5645–5657.

Aditya Wiha Pradana and Mardhiya Hayaty. 2019. The effect of stemming and removal of stopwords on the accuracy of sentiment analysis on indonesian-language texts. *Kinetik (Malang)*, pages 375–380.

David Ramamonjisoa. 2014. Topic modeling on users's comments. In *2014 Third ICT International Student Project Conference (ICT-ISPC)*, pages 177–180. IEEE.

Appendix

Topic	**Weights**	**Color**
Cancer	3647	violet
Coronavirus	2902	indigo
Women's health	2128	blue
Emergency health conditions	1979	green
US Presidency	461	brown
Blood donation	1107	orange
Drug tests	3577	red
US & Coronavirus	1726	cyan
Cardiac disease	1659	pink
Disease treatment	1905	darkgreen

Table 3: The topics and their associated weights each assigned a distinct color for Propagation graph (I)

Topic 1	**Topic 2**	**Weights**	**Color**
Cancer	Drug tests	4173	violet
Cancer	Coronavirus	2175	violet
Women's health	US Presidency	47	blue

Table 4: The topics and their associated weights each assigned a distinct color for Propagation graph (II)

Problem Statement	Aditya and Anwar
Related Work	Bithy, Hasib, and Mohammed
Topic Modeling Methodology and Performance Metrics Calculations	Aditya, Anwar, and Hasib
Propagation Graphs	Aditya, Bithy, and Mohammed
Project Approach, Results, and Final Report	All members

Table 5: Author Contributions