2022 International Conference on Trends in Quantum Computing and Emerging Business Technologies (TQCEBT)
CHRIST (Deemed to be University), Pune Lavasa Campus, India. Oct 14-15, 2022

2022 International Conference on Trends in Quantum Computing and Emerging Business Technologies (TQCEBT) | 978-1-6654-5361-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/TQCEBT54229.2022.10041450

Topic Modelling of ongoing conflict between Russia and Ukraine

Pradhan Nayak
Liverpool John Moores University
Liverpool, UK
pradhan.nayak3@gmail.com

Lakshmi J V N
Patel Institute of Science and Management, Sunstone
Bangalore, India
Lakshmi.jvn@sunstone.in
ORCID ID - 0000-0001-9230-2521

Vandana V. Bhagat
CHRIST (Deemed to be University),
India
vandana.bhagat@christuniversity.in
line 5: 0000-0003-2106-6298

***Abstract*— Online news sites provide hotspots to extract popular ratings and opinions on a wide range of topics. Realizing what individuals are referring to and understanding their concerns and suppositions is exceptionally significant to organizations and political missions. Furthermore, it is incredibly difficult to physically peruse such enormous volumes of data and gather the themes. Keeping in mind the prevailing plight of war-torn nations such as the recent conflict between Russia and Ukraine. This study performs aims to perform topic modelling using LDA (Latent Dirichlet Allocation) and text analysis on datasets collected from various online news websites. To increase the accuracy and efficacy of the topic modelling, a comparative analysis is proposed that elevates the performance of machine learning models. This study also develops an algorithm where the entire process can be automated from the point of data collection to finding optimum array of topics in the given dataset. Searching for insights from the collected information can therefore become very tedious and time-consuming. Topic modelling was designed as a tool to organize, search, and understand vast quantities of textual information. The topic model using LDA was utilized to do a text analysis for this research. In the beginning, researchers have scraped a total of 1178 articles that covered the war conflict between Russia and Ukraine from December 1, 2021, to May 16, 2022. After that, researcher built the LDA model and modified hyper parameters based on the coherence score Cv that was used for the model evaluation technique. When using the most effective model, prominent topics, and representative documents pertaining to each topic, topic allocation among the documents, and potential enhancements are covered in the last section.**

Keywords—LDA, Topic modelling, Russia, Ukraine, Topic.

I. INTRODUCTION

Relations between the two countries Ukraine and Russia are presently at odds: the Russia-Ukraine conflict erupted following Russia's annexation of Crimea from Ukraine in the year 2014. The Ukrainian Constitution was amended in 2019, protective the country's diplomatic course towards international organization members and European nations as irreparable. Throughout 2021, Russia's military presence close to Ukraine's border has inflated tensions and worsened bilateral ties, with the America causation out a robust message that any intrusion would be met with devastating economic sanctions for Russia. These activities are still in progress in 2022.

A mathematical technique that uses an unsupervised machine learning approach to automatically identify the content and subjects covered by a huge collection of news items. This technique is used to extract relevant information from online news sources. The approach that is being described here is called topic modelling, and it is one of the solutions that can be found for topic search and topic analysis. The algorithm known as Latent Dirichlet Assignment (LDA) has received the greatest attention in terms of research. This is a generative statistical Bayesian algorithm. The system works on the assumption that the document has a limited number of subjects, each of which is comprised of a group of words [1].

LDA has recently become popular in text mining studies in a variety of fields, including natural language processing, extracting data, sentiment analysis, trending analytics and literature research. This technique has also been used as a reasonable method in some research that analyzed online consumer reviews in a variety of industries. Topic models, which were originally developed to analyze textual data, are currently being used to a variety of data sources, including social media text data, image data, video data, and genetical data. For these reasons, LDA was chosen as the topic modelling technique in this study [2].

In the earlier researchers started employing topic modelling as a means for developing more precise information retrieval systems. This was accomplished through the use of topic modelling. Latent Semantic Analysis (LSA) is the acronym that was given to the technique that was developed by Deer wester, Dumais, Furnas, Landauer, and Harshman. These individuals are credited with the creation of an early topic modelling technique. Previously, users were required to use the same words in their searches. Because this approach represents an improvement over other methods of document retrieval, this was something that was previously impossible to achieve.

Other introductions to LDA take a more metaphorical approach to attempting to explain the process, in contrast to the numerous presentations that Centre on the mathematical aspects of topic modelling and LDA. The idea behind LDA is based on the assumption that each document contains a variety of themes, and that each of those topics is denoted by a distinct group of words. Assuming this to be the case, LDA looks for groups of words that appear frequently together and considers them to be subjects. The observable words that actually appear in the papers can be used as a measurement tool to determine how widely distributed these subjects are across the texts.

II. RELATED WORK

A. Topic Modelling

Since its introduction, the topic model has been the standard method utilized in a variety of domains, including

but not limited to those of topic detection, multidocument summary, word sense recognition and disambiguation, emotional analysis, and information retrieval, amongst others. These fields obtain topics through training with the topic model. It has been emphasized, with the goal of facilitating users' comprehension of the subject matter at hand, the importance of selecting an appropriate vocabulary to convey a sense of the subject matter [3].

The process of discovering latent semantic patterns, also known as topics, in a collection of unstructured documents is referred to as topic modelling [2].Topic modelling is a probabilistic method. The process of characterizing the semantic structure of documents in accordance with the themes that were discovered is known as topic modelling [4]. This strategy is predicated on the idea that subjects have a certain proportion of a document's total word count that can be attributed to a random probability distribution [5]. Because a document is naturally connected to a certain subject, it is reasonable to anticipate that particular phrases will appear more frequently within the body of the document. The topics that are uncovered by using topic modelling are in fact semantic clusters that are formed when words that are frequently used together in a document come together. During the process of identifying latent semantic structures, the probability distribution of each subject, the distribution of topics per document, and the distribution of topic assignments per word in each document are all calculated [6].

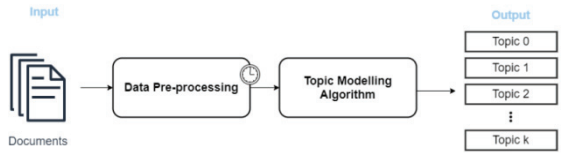

Fig. 1. Topic modelling process, [20]

B. *Latent Dirichlet allocation*

Latent Dirichlet allocation, often known as LDA, is a method of unsupervised topic modelling that is widely used to find the range of topics that are present in text corpora and to unearth any underlying trends. In general, modelling text corpora is a difficult task because its goal is to identify brief descriptions of the documents contained in a collection.

This is done in order to "enable efficient processing of large collections while preserving the essential statistical relationships that are useful for basic tasks such as classification, novelty detection, summarization, and similarity and relevance judgments." [4].

By utilising mixture models, which are based on the assumption that words within a document and the sequence of the documents included within a corpus may be interchanged, the LDA makes it possible to capture the statistical structure contained inside a text. The papers themselves are considered to be "random mixes over latent subjects" in the context of this generative and probabilistic model. An LDA algorithm is used to extract the "set of K frequencies that minimise the reconstruction error of the original documents" when it is given a group of documents that are "represented as a vector of word frequencies" and a predetermined number of K topics. In a nutshell, this process begins with a collection of documents that are "represented as a vector of word frequencies" The challenge of model selection presents itself when deciding on the number of latent themes to be estimated by the LDA, which is denoted by the letter K [7].

Before beginning to initialise the model, it is necessary to discover this unidentified number. The LDA is featured here because to the fact that it is both effective and widely used. It is a well-known technique for modelling themes in text data, and the end result is a weighted list of topics that applies to every article in the corpus. Each individual subject is made up of a probability distribution that specifies the possibility of a certain word appearing in a particular topic.

This distribution serves as the basis for producing a representation of the text. Based on Bayesian statistical models, LDA "is a generative probabilistic model of a corpus." This model is used to analyse large amounts of text.

The LDA model is capable of both generation and deduction of new information. Our latter capacity will be the major focus of this investigation, which will derive the precise subjects which are hidden inside each of the two datasets.

Since its inception in 2003, several iterations of the LDA method have been developed, despite the fact that the traditional form of the algorithm continues to be one of the most well-known and widely implemented topic modelling techniques.

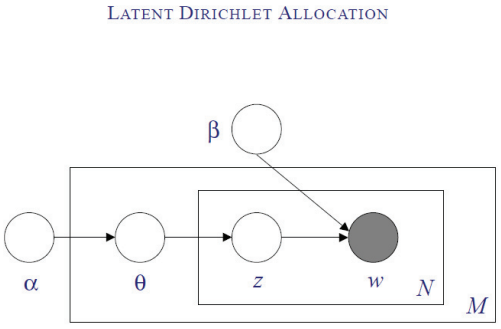

Fig. 2. Graphical representation of LDA model [4]

C. *Gensim LDA Model*

Genism, which is a set of tools for natural language processing, comes equipped with a practical implementation of the LDA paradigm [8] [9]. Adjustments may be made to the number of subjects, iteration, chunk size, and passes inside the genism LDA model. These are the four primary parameters that can be modified. The number of topics is the most crucial parameter in a topic model because having too few topics will cause the model to be inaccurate (also known as underfitting), whereas having too many subjects would cause the model to be overfit. The term "iteration" refers to the maximum number of times each document in the corpus must be iterated through before the probability of every subject can be calculated. The term "chunk size" refers to the amount of documents that are contained inside each individual training chunk since the model is trained on one data chunk at a time. The number of times that the entire corpus would be trained on is referred to as the passes. If the iterations and passes were made larger, the model would become more accurate; however, this would result in the training process taking significantly more time.

Therefore, in order to compensate for this, we will select the fewest possible options that are precise enough [10].

D. MALLET LDA Model

[1] For the purpose of topic modelling, we made use of the latent Dirichlet allocation that uses the MALLET tool that is based on Java in addition to scikit-learn and Gensim Python packages. The Gensim and sci-kit learn implementations, respectively, make use of the online nonlinear Bayes method in their respective processes. The MALLET implementation makes use of the Gibbs sampling process in order to collect data. The Gibbs sampling methodology is seen to be to be more objective and then the online nonlinear Bayes method, which enables a faster analysis of the data at the risk of perhaps being biased, but the Gibbs sampling approach is believed to be more accurate overall. On the other hand, the online nonlinear Bayes technique enables a more rapid examination of the data; however, it does so at the price of calling for an increased amount of system resources. MALLET was picked to serve as the final implementation technique moving forward as a result of its effectiveness in a broad number of sectors. This decision was made in light of the fact that MALLET has been used successfully. Because we wanted as many as we could get away with without compromising the overall quality of the models or making the topics themselves too detailed, we came to the conclusion that twenty was the optimal number of subjects to cover.

In order to determine which of the possible starting settings for MALLET would produce the best results, some more fine tuning was carried out. In the end, we came to the conclusion that we should select the model that had the highest coherence score, which we ascertained by assembling a last group of potential candidate models.[11].

III. Research Methodology

The purpose of this research is to follow a technique that can be employed in any context of an online news corpus and might potentially be applied to any online text content. The approach presented in this study is a framework that is built on well-known procedures for cleaning and processing text data. In order to validate the methodology, the researchers had employed a case study in the form of an online news article. The approach produces a list of subjects into which news articles are divided. The framework is judged justified if the combination of internet news articles makes sense and is logical, and can be examined by a knowledgeable individual in this research subject. Unlike other strategies, such as supervised learning, the technique of assessing effectiveness does not yield the desired results. A skilled expert in the field, on the other hand, will most likely be able to spot an inaccurate result. Topic modeling is used for efficiency rather than precision[12].

A. Scraping the web data

The suggested research focuses on two areas: I) building and implementing a data mining technique to extract conflict information from online media, and ii) processing and analysing the online news data using machine learning techniques. One of the aims of this study is to design a focused python script that follows a set of well-defined criteria for visiting online pages and gathering data without breaking the data scraping rules of each web page, assuring ethics. To do this, we're building a dedicated web scrapper. The web scrapper will explore each website until it finds an article. When an article is located, its content is tentatively processed before being downloaded. Relevant articles are downloaded [13].

After carefully looking through many news reports about war, it's clear that the Russia, Ukraine keywords are commonly utilised in web publications. This collection of keywords is updated so a string-matching algorithm may be performed with the article's text to determine if it's about a battle.

When provided a URL, this module will respond with right or wrong to indicate data retrieval authorization. If a URL doesn't allow scraping, the scrapper won't investigate it until it finds an article.

Few URLs for articles end with page or news story numbers. Such URLs may lead to uncertainty and needless investigation of the content. The scraper focuses on digging further through URLs is found.

After finding article URLs, analyse each report's text and save it for further review. Newspaper library is one technique for processing web page content. This library scrapes article data including title, content, date, source url, and more. After being processed and briefly saved, articles are examined for their substance.

B. Data selection

This research utilized the dataset that is freely accessible on a news website. We began collecting data in December 2021, when news came that Russia was planning a military exercise near the Ukraine border, as well as a possible invasion. We initially collected only 30 publications during the research proposal stage, but after further investigation, we discovered that this data is insufficient to make any serious assessment.

Then, from December 2021 to May 16th 2022, we collected web news pieces once a month. In the csv file, the extracted URL was saved. The pieces mostly consisted of professional viewpoints, journalistic reports, events during the war, and on-the-ground witnesses.

The news article URLs recorded in a csv file could not be used directly for analysis; instead, the data for the news items was scraped from the URLs. The scraped data was then saved in a separate csv file with the title of the article in one column and the text of the article in another.

A total of 1178 news pieces were gathered from various internet news sources. We utilized python programming and the nltk libraries to scrape the data. During the extraction, we discovered that some of the URLs were broken or that scraping the data was restricted. As a result, we modified the current python code to extract just the links that were successful in bypassing the website links and added a sleep duration of 5 seconds because there were so many links that were related to the same website. Finally, the extracted data was saved in a separate csv file, which will be utilized for data preprocessing and modeling in the future

C. Exploratory Data Analysis

Exploratory data analysis, or EDA, is a technique used in statistics that helps data analysts identify the primary information in the data, primarily via the use of visual approaches. This understanding may then lead to the

development of hypotheses and the implementation of additional experiments. The exploratory data analysis may be broken down into two categories. In the first place, the approach might either be graphical or non-graphical. And second, the approach may be univariate or it could be multivariate. Following the completion of EDA, word cloud visualization was utilized in order to comprehend the phrases pertaining to the ongoing conflict. [14].

D. Data Preprocessing

To convert the information into a system-readable vector shape, pre-processing is required.

- Sentence Tokenization: This is a technique for segmenting text into sentences. To split up the sentences, the NLTK library python's sentence tokenizer is used.
- Remove Special Characters: It's possible that the textual material contains a few characters that aren't needed. All of the unnecessary characters had been removed.
- Word Tokenization: Using phrase spaces, each of the object's sentences has been broken up into phrases.
- Elimination of stop phrases: Stop phrases are sentences that are designed to go unnoticed while the text is being processed. The entire body of text was combed through, and any instance of a phrase that may be construed as a stop phrase was deleted.
- Duplicate Word Removal: Words from each phrase that appeared more than once have been removed, with the exception of the first occurrence of each word, which has been preserved.
- Lemmatization: This is a method for determining the root of each sentence. The phrases in the textual content have all been lemmatized.

Text pre-processing tools, such as the Gensim package in Python, which includes "Gensim. Corpora", "Gensim.utils", and "Gensim models," have been utilized once the data have been imported. Text normalization was started once the text was collected. This involved converting all letters to lowercase, turning numerals into words, removing punctuation and white space, and eliminating abbreviations and stop words. A further split into tokens is performed on the provided text by utilizing the Natural Language Toolkit (NLTK) package.

After the input has been pre-processed into tokens, the Phrases () module in Gensim is used to turn the tokens into bigrams and trigrams. For the purposes of modelling, bigrams are defined by the frequency at which two words occur together often. Lemmatization is a technique that involves converting various forms of the word back into its base form. Only the "Noun," "Adj," "Verb," and "Adverb" parts of speech are included in the definition of the function. This is because including any additional parts of speech would not contribute to the creation of a decent topic model and would lead to a low coherence score [15].

E. Perplexity Score

The model that contains the fewest amount of complexities should be chosen because this will help establish the optimal number of themes. When the perplexity is lower, it indicates that the model has a stronger ability to predict how new text will behave. At the same time, the model's ability to generalize increases in proportion to the degree to which the ambiguity has been reduced. The perplexity formula [16] is as follows:

$$perplexity\,(D) = exp\left(-\frac{\sum_{d=1}^{M} log\ p(w_d)}{\sum_{d=1}^{M} N_d}\right)$$

Fig. 3. The formula for perplexity [18]

D represents the test set in the corpus, and the test set has a total of M documents. Nd represents the number of words in each document d, wd represents the word in document d, p(wd) is the probability of the word wd in the document.

F. Coherence Measures

Our brand-new unifying framework depicts the coherence measure as the assembly of freely combinable components. This is how the measure may be understood. As a direct consequence of this, it will be possible to establish measures that currently exist in addition to measures that have not been studied. The components are arranged in dimensions that cover the entirety of the configuration space that is employed in the coherence measurement. Each dimension may be seen as a distinct collection of movable and interchangeable components. The degree to which individual words or subsets of the words in a set of words hang together and fit together is what is supposed to be assessed by coherence in the set of words. Therefore, the first dimension is the type of segmentation that is used when breaking a word setup into more manageable pieces. This is because the word setup is the starting point for the segmentation. When the text is broken up into word pairs, for example, these pieces are juxtaposed against one another to highlight their differences. The many methods of segmentation are grouped together and denoted by the letter S in this diagram. The confirmation measure is the second dimension, and it analyses how well a given pair, such as the NPMI of two words, agrees with one another. An example of this would be comparing the NPMI of two different terms. The letter M denotes the total number of confirmation measures. Word probabilities, which may be assessed in a wide variety of various ways, create the third dimension of the configuration space, which is called the word probability space. These probabilities are used in the confirmation measures. P refers to the collection of approaches that may be used to estimate word probabilities. The fourth dimension is made up of the many techniques that may be used to aggregate scalar values that are computed by the confirmation measure. denotes the collection of aggregation functions. Figure 1 presents an overview of our system's process, which can be broken down into four distinct stages. At first, the word set t is cut up into a number of different subsets of pairs of words called S. Second, a word's probability, denoted by the letter P, is determined by using a specific reference corpus. The confirmation measure uses both the set of word subsets S and the computed probabilities P in order to determine the agreements of pairs of S. Both of these are utilized by the confirmation measure. In the end, all of these values are combined into a single coherence value denoted by C [17].

G. Pointwise Mutual Information(PMI)

The methodology that prioritizes the real phrases over the other terms is an example of an efficient approach to semantic phrase extraction. Considered in this study is the pointwise mutual information (PMI), which is the mutual information-based measure that has seen the most widespread application. PMI was introduced by Church and Hanks(1990)(Kherwa and Bansal, n.d.). PMI is a measure of how much the actual probability of a particular co-occurrence of events P (X, Y) differs from what we would expect it to be on the basis of the probabilities of the individual events and the assumption of independence P(X), P(Y)in some situation. In PMI, low-frequency events receive high scores, which indicates that rare word pairings have a tendency to dominate frequent events. Therefore, the less common occurrences come first in the ranking. This indicates that the PMI of completely connected events words will be higher when the combo will be occurring less frequently. Therefore, the PMI might be interpreted as a measure of independence rather than as a measure of correlation, as the previous sentence stated. Independence rather than as a measure of correlation.

$$I(W_1, W_2) = \log_2 \frac{P(W_1, W_2)}{P(W_1).P(W_2)}$$

Fig. 4. Figure 3. 10. Interpretation of PMI scoring [6]

H. Evaluating Hyperparameter

There are four hyperparameters to set when using the Gibbs sampling generative model for LDA:

- The number of topics K to generate from the corpus;
- The hyperparameter α, which influences the distribution of the topics per document. Smaller α values lead to fewer topics per document.
- The hyperparameter β, which influences the term distribution in each topic. Smaller β values lead to topics with fewer words.
- The number of Gibbs iterations N; this parameter is specific to the Gibbs sampling generative model.

I. pyLDAvis to visualize topic relationships

Access to two crucial pieces of information may be gained through the use of the pyLDAvis package in Python. The circles, each representing a distinct concept, are shown below. The tight closeness of the circles is a helpful illustration of how closely connected the topics are. The distances between the probability distributions of each topic are used to determine how they are mapped into a two-dimensional space using "dimensionality reduction" techniques such as "principal component analysis" and "t-sne". The results of this study show whether or not our model generates distinct subjects. The model parameters needs to be adjusted and expand the number of subjects such that the overlapping of the circles is reduced to the greatest extent feasible.

J. Tune relevancy score to prioritize terms more exclusive to a topic

Because they are used so frequently throughout a corpus, it is likely for words that reflect a certain topic to get a high score solely because of how frequently such words are used. The relevance score assists in selecting phrases that correspond more specifically to a certain topic, which in turn serves to make that topic more understandable [19].

The definition of the relevance of a term w to a topic k is as follows: where _kw is the likelihood of the word w appearing in the topic k, and where _kw/p kw is the lift in the term's likelihood within a topic relative to its marginal likelihood throughout the whole corpus (this helps discards globally frequent terms). A lower value for k confers greater significance on the second term (kw/pk), which in turn confers greater significance on the notion of topic exclusivity. Once more, pyLDAvis is a tool that we may utilise for this. On changing the value form from 1 to 0.6, Lambda must be adjusted in order to produce the outcome that makes the most intuitive sense, and the output must be obtained using the lambda value that is ideal for the situation.

IV. Results And Discussions

This chapter provides an in-depth analysis of the results obtained from the training of the various LDA Topic models. The results of the created models have been shown through the use of pyLDAvis, which demonstrates the topic clustering and prominent keyword results obtained from LDA models. Metrics like the perplexity score and the coherence score have been offered as a way to demonstrate how well the model is able to distinguish between different types of topics. Tabular form has been used to offer a comparison of the LDA Gensim model with the LDA MALLET model for this study.

The research have gone through a few of the hyper-parameters that were utilised during the modelling process. A number of different iterations of the model have demonstrated the change in the total number of articles as well as the total number of topics. The results of running Gensim and MALLLET through the LDA model are shown here.

A. LDA Model outcome using Gensim and MALLET method

In the third baseline model, there were 1178 pieces of text data from publications that were done. exploratory data analysis of the dataset that was extracted, text data preprocessing such as tokenizing the entire text data and then building bigrams and trigrams containing only nouns and verbs, removing stopwords, taking into consideration only English words, removing punctuation, and lemmatizing the text utilizing Gensim preprocess and the spacy library. Then, utilizing genism and MALLET, the study established the corpus and the dictionary that are prerequisites for constructing the LDA model.

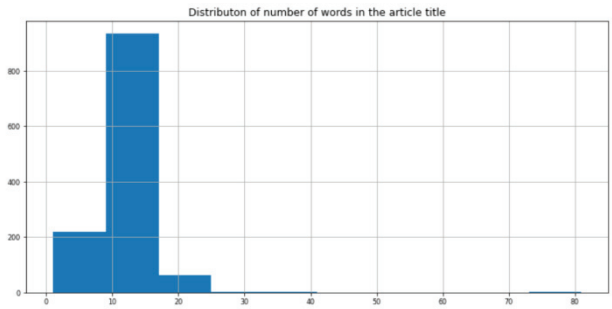

Fig. 5. Third Baseline Model –distribution of the words in the news article title

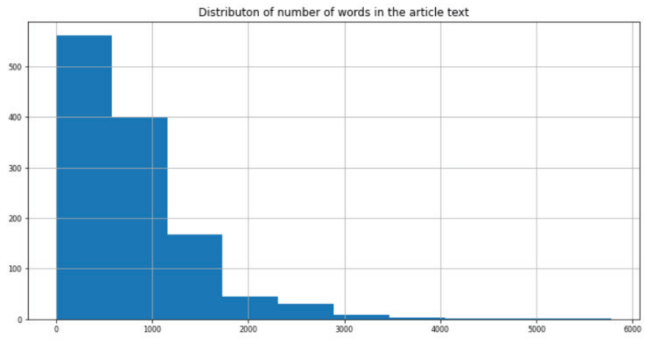

Fig. 6. Third Baseline Model –distribution of the words in the news article text

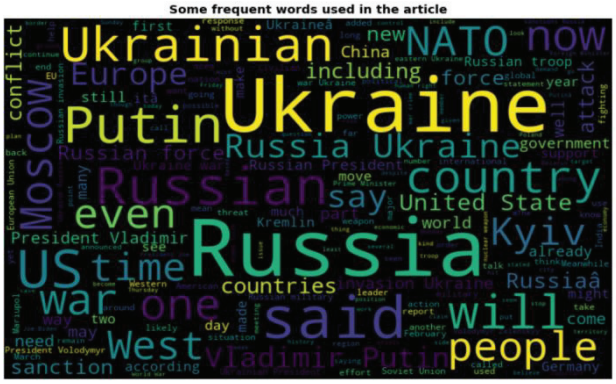

Fig. 7. Third Baseline Model –Frequent words used in the article text

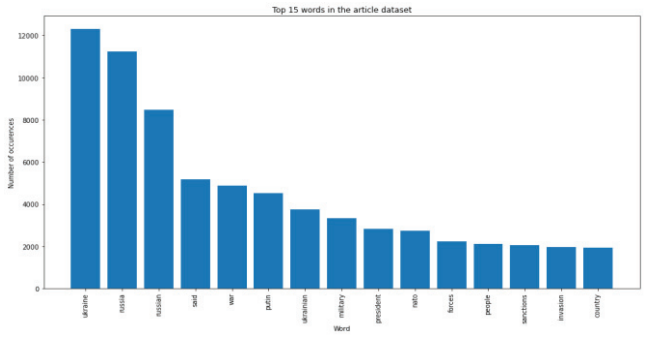

Fig. 8. Third Baseline Model –Top 15 words used in the article

This model was run considering a range of topic between 5 and 25 and check the coherence score. whichever gave the highest coherence score that topic was considered the best topic number for the given dataset.

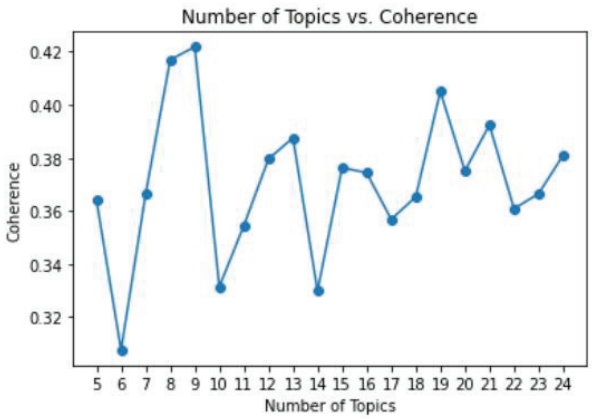

Fig. 9. Third Baseline Model – shows number of topics vs coherence score using Gensim

The results for the third base model can be visualized through pyLDAvis and are presented below

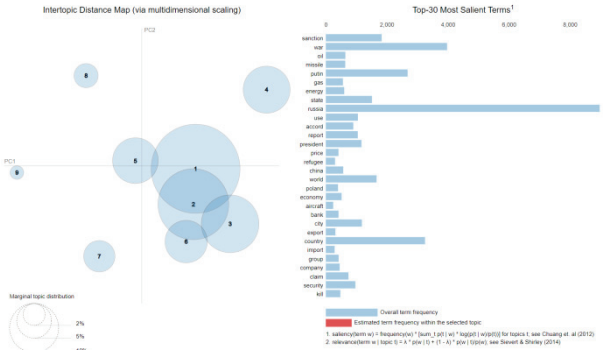

Fig. 10. . Third Baseline Model – showing topic clustering and 30 most salient words in the entire dataset.

TABLE I. DISPLAYING RESULTS FOR THE TOP 9 TOPICS FROM THE THIRD BASE MODEL.

	0	1	2	3	4	5	6	7	8	9
Topic 1	russia	force	war	country	city	attack	kyiv	troop	day	border
Topic 2	russia	putin	war	sanction	country	force	world	invasion	time	use
Topic 3	russia	putin	china	security	country	president	world	europe	war	nato
Topic 4	russia	sanction	oil	gas	energy	country	price	economy	export	bank
Topic 5	russia	war	state	food	use	world	country	city	report	government
Topic 6	war	russia	world	country	thing	power	putin	way	conflict	time
Topic 7	war_crime	war	crime	bucha	court	body	evidence	genocide	commit	russia
Topic 8	refugee	missile	aircraft	accord	poland	report	country	state	help	destroy
Topic 9	symbol	state	use	president	letter	shield	patron	group	order	israel

Calculating the coherence score using LDA MALLET method

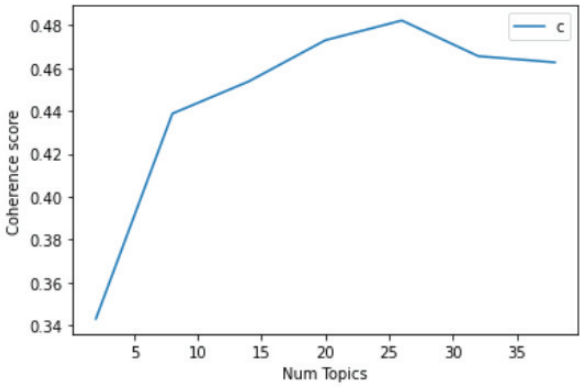

Fig. 11. Third Baseline Model – shows number of topics vs coherence score using MALLET

It can be observed that results have improved drastically the coherence score is peak for number of topics equal to 27 and then the coherence has a slight dip and remains throughout for more than 35 topics.

TABLE II. DISPLAYS10 MOST DOMINANT TOPICS FOR THE THIRD BASE MODEL

	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text
0	0	9.0	0.1745	putin, russia, soviet_union, year, power, lead...	[week, official, statement, tension, russia, h...
1	1	9.0	0.1842	putin, russia, soviet_union, year, power, lead...	[crisis, create, russia, expansion, europe, mo...
2	2	8.0	0.1997	russia, border, troop, moscow, force, invasion...	[president, putin, admit, member, threat, role...
3	3	8.0	0.1387	russia, border, troop, moscow, force, invasion...	[reservist, force, exercise, outskirt, kyiv, k...
4	4	11.0	0.0658	thing, happen, kind, point, lot, war, end, que...	[president, putin, centre, world, attention, w...
5	5	9.0	0.3730	putin, russia, soviet_union, year, power, lead...	[russian_troop, mass, border, worry, invasion,...
6	6	12.0	0.2188	putin, russia, threat, war, world, country, at...	[surprise, shock, force, secure, speed, russia...
7	7	8.0	0.2368	russia, border, troop, moscow, force, invasion...	[moscow, russia, warn, measure, reject, securi...
8	8	3.0	0.1798	medium, russia, service, company, state, infor...	[government, sunday, mastermind, attack, targe...
9	9	3.0	0.1597	medium, russia, service, company, state, infor...	[evidence, russia, cyber, attack, knock, gover...

The most representative document for each topic for the third base model is presented below.

TABLE III. DISPLAYS10 MOST REPRESENTATIVE DOCUMENT FOR EACH TOPIC FOR THIRD BASE MODEL

	Dominant_Topic	Topic_Keywords	Num_Documents	Perc_Documents
0.0	9.0	putin, russia, soviet_union, year, power, lead...	63.0	0.0520
1.0	9.0	putin, russia, soviet_union, year, power, lead...	28.0	0.0231
2.0	8.0	russia, border, troop, moscow, force, invasion...	42.0	0.0347
3.0	8.0	russia, border, troop, moscow, force, invasion...	45.0	0.0372
4.0	11.0	thing, happen, kind, point, lot, war, end, que...	43.0	0.0355
5.0	9.0	putin, russia, soviet_union, year, power, lead...	34.0	0.0281
6.0	12.0	putin, russia, threat, war, world, country, at...	32.0	0.0264
7.0	8.0	russia, border, troop, moscow, force, invasion...	32.0	0.0264
8.0	3.0	medium, russia, service, company, state, infor...	75.0	0.0619
9.0	3.0	medium, russia, service, company, state, infor...	53.0	0.0438
10.0	13.0	war, force, fight, russia, army, russians, tim...	27.0	0.0223
11.0	2.0	week, official, invasion, poland, plan, day, d...	22.0	0.0182
12.0	8.0	russia, border, troop, moscow, force, invasion...	35.0	0.0289
13.0	16.0	russia, china, germany, europe, crisis, beijin...	42.0	0.0347
14.0	16.0	russia, china, germany, europe, crisis, beijin...	44.0	0.0363
15.0	16.0	russia, china, germany, europe, crisis, beijin...	36.0	0.0297
16.0	16.0	russia, china, germany, europe, crisis, beijin...	58.0	0.0479
17.0	13.0	war, force, fight, russia, army, russians, tim...	81.0	0.0669
18.0	24.0	country, security, nato, alliance, finland, ru...	51.0	0.0421
19.0	24.0	country, security, nato, alliance, finland, ru...	67.0	0.0553

Topic distribution across documents for the third base model is shown below.

TABLE IV. DISPLAYS TOPIC DISTRIBUTION ACROSS DOCUMENTS FOR THIRD BASE MODEL

	Dominant_Topic	Topic_Keywords	Num_Documents	Perc_Documents
0.0	9.0	putin, russia, soviet_union, year, power, lead...	63.0	0.0520
1.0	9.0	putin, russia, soviet_union, year, power, lead...	28.0	0.0231
2.0	8.0	russia, border, troop, moscow, force, invasion...	42.0	0.0347
3.0	8.0	russia, border, troop, moscow, force, invasion...	45.0	0.0372
4.0	11.0	thing, happen, kind, point, lot, war, end, que...	43.0	0.0355
5.0	9.0	putin, russia, soviet_union, year, power, lead...	34.0	0.0281
6.0	12.0	putin, russia, threat, war, world, country, at...	32.0	0.0264
7.0	8.0	russia, border, troop, moscow, force, invasion...	32.0	0.0264
8.0	3.0	medium, russia, service, company, state, infor...	75.0	0.0619
9.0	3.0	medium, russia, service, company, state, infor...	53.0	0.0438
10.0	13.0	war, force, fight, russia, army, russians, tim...	27.0	0.0223
11.0	2.0	week, official, invasion, poland, plan, day, d...	22.0	0.0182
12.0	8.0	russia, border, troop, moscow, force, invasion...	35.0	0.0289
13.0	16.0	russia, china, germany, europe, crisis, beijin...	42.0	0.0347
14.0	16.0	russia, china, germany, europe, crisis, beijin...	44.0	0.0363
15.0	16.0	russia, china, germany, europe, crisis, beijin...	36.0	0.0297
16.0	16.0	russia, china, germany, europe, crisis, beijin...	58.0	0.0479
17.0	13.0	war, force, fight, russia, army, russians, tim...	81.0	0.0669
18.0	24.0	country, security, nato, alliance, finland, ru...	51.0	0.0421
19.0	24.0	country, security, nato, alliance, finland, ru...	67.0	0.0553
20.0	15.0	russia, india, vote, resolution, moscow, situa...	27.0	0.0223
21.0	1.0	president, biden, leader, administration, bide...	33.0	0.0273
22.0	8.0	russia, border, troop, moscow, force, invasion...	70.0	0.0578
23.0	12.0	putin, russia, threat, war, world, country, at...	66.0	0.0545
24.0	8.0	russia, border, troop, moscow, force, invasion...	62.0	0.0512
25.0	2.0	week, official, invasion, poland, plan, day, d...	43.0	0.0355

Word Clouds of Top N Keywords in Each Topic for the third base model is shown below.

Fig. 12. Third Baseline Model – word cloud for top 4 topics in the corpus

B. *Comparison between Models*

The online news article on the Russia-Ukraine conflict dataset is used in this research, and we built topic models, one based on LDA Gensim, which gave us relatively good topic models but the coherence score of 0.42 kept varying between the number of topics, and in comparison, the LDA with MALLET implementation gives a better coherence score of 0.48 and remains consistent as the number of topics increases.

C. *Summary*

A comparison between the LDA model with Gensim and LDA model with MALLET topic models gave a fair understanding of the coherence score obtained by data pre-processing and tuning hyper-parameters. The final selected model, Topic Model with LDA MALLET established better results for the online news article of Russia Ukraine conflict dataset. We saw that the coherence score was maximum with value 0.48 when the number of Topic model was 25. We also saw the dominant topics within the corpus listed out.

V. FUTURE RECOMMENDATIONS

In spite of the fact that topic modelling has been examined to a significant degree in a number of disciplines of inquiry, our investigations have revealed that certain sectors in which topic modelling may be applied have not yet been thoroughly investigated. Processing speech and images, as well as locating recurring themes in audio and video data sets, are examples of the areas that fall under this category. It is essential to identify specific features via which topic modelling may be comprehended in a more fundamental manner. In this article, several difficult aspects of topic modelling, such as the visualisation of topics, the interpretation of topics as people are covered. Models, in addition to the Stability of Topic Modelling, are uncovered and investigated in this process. In spite of the substantial advancements that have been made in the field of topic modelling over the course of the past two decades, a great number of questions have not been resolved. Among these are: deciding how many different topics to put in the group from which to select one. Although there are a variety of methodologies, such as log-likelihood, harmonic mean, perplexity and cross validation the results of these and other methods are still not entirely obvious.

There is presently no study available that has been conducted in great depth on the topic of previous selection. Which inference method is best suited for the subject matter being discussed? For the past two decades, the discussion around modelling has been mostly centred on nonlinear inference and Gibbs sampling; yet, the question of whether approach is more effective has not been satisfactorily resolved.

Therefore, as a result of the growth in topic modelling, numerous expanded versions of LDA that are based on related theory have evolved and been utilised in diverse applications throughout the course of the previous two decades. These variants of LDA have been employed.

It has been demonstrated throughout this piece of writing that the LDA topic models can exhibit a high degree of variability from one run to the next. It was further explored whether or not the impact of increasing the number of passes, which enhanced the model-to-model stability, was significant. On the other hand, utilising Ensemble LDA may result in a major improvement of the output in its ultimate form. This was achieved by significantly cutting down on the amount of overlap as well as the quantity of noise. Ensemble LDA is famous not just for the fact that it resolves the issue of instability, but also for the fact that it enables users to pick the topic size that functions most effectively for the model. Both of these features contribute to Ensemble LDA's popularity.

VI. Discussion and Conclusion

The first baseline model that has been chosen had a very small number of articles, which meant that LDA was unable to come up with useful topic segregation and was only able to see three clusters of topics. These clusters of topics are distinct, but they are not enough to understand what exactly happened and what was said during the conflict and the implications on the global economy.

Then, the second baseline model was run, taking into consideration a range of topics this time with 1062 news articles. Some improved data pre-processing has been done from the first model, and as a result, it is noticed that a significant increase in the coherence score value, which was 0.45 for LDA with Gensim and 0.47 for LDA with MALLET implementation. Additionally, improvement in both the topic segregation and the salient keyword quality has been done.

The third and final baseline model was run with 1178 articles, which included the latest collection of news articles up until May 16th, 2022. Here, a slight decrease in coherence score for LDA Gensim was observed, which demonstrates that there is some inconsistency in this model when it comes to coherence score. This is the case despite the fact that the baseline model 2's data pre-processing and hyperparameters were used. The maximum coherence score in our LDA MALLET implementation was 0.48, but we did notice a modest increase in this area. In light of this, we are able to draw the conclusion that LDA with MALLET implementation produces better topics for a particular corpus, and that the coherence score appears to be better and more consistent, despite the fact that the number of topics was increased, in comparison to LDA with Gensim.

References

[1] Porturas, T. and Taylor, R.A., (2021) Forty years of emergency medicine research: Uncovering research themes and trends through topic modeling. American Journal of Emergency Medicine, [online] 45, pp.213–220. Available at: https://doi.org/10.1016/j.ajem.2020.08.036.

[2] Gurcan, F. and Cagiltay, N.E., (2019) Big Data Software Engineering: Analysis of Knowledge Domains and Skill Sets Using LDA-Based Topic Modeling. IEEE Access, 7, pp.82541–82552.

[3] Han, W., Tian, Z., Zhu, C., Huang, Z., Jia, Y. and Guizani, M., (2021) A Topic Representation Model for Online Social Networks Based on Hybrid Human-Artificial Intelligence. IEEE Transactions on Computational Social Systems, 81, pp.191–200.

[4] Blei, D.M., Ng, A.Y. and Jordan, M.I., (2003) Latent Dirichlet allocation. Journal of Machine Learning Research, 34–5, pp.993–1022.

[5] Bawden, D. and Robinson, L., (2009) The dark side of information: Overload, anxiety and other paradoxes and pathologies. Journal of Information Science, 352, pp.180–191.

[6] Kherwa, P. and Bansal, P., (n.d.) Semantic N-Gram Topic Modeling. 726.

[7] Roque, C., Lourenço Cardoso, J., Connell, T., Schermers, G. and Weber, R., (2019) Topic analysis of Road safety inspections using latent dirichlet allocation: A case study of roadside safety in Irish main roads. Accident Analysis and Prevention, [online] 131August, pp.336–349. Available at: https://doi.org/10.1016/j.aap.2019.07.021.

[8] WANG Minghao, D.P.M., (2021) How Pandemic Spread in News: Text Analysis Using Topic Model WANG Minghao , Dr. Paolo MENGONI , IEEE Member. IEEE.

[9] Jones, S.M., Weigle, M.C. and Nelson, M.L., (2018) The Off-Topic Memento Toolkit. [online] Available at: http://arxiv.org/abs/1806.06870.

[10] Samin, H. and Azim, T., (2019) Knowledge Based Recommender System for Academia Using Machine Learning: A Case Study on Higher Education Landscape of Pakistan. IEEE Access, 7, pp.67081–67093.

[11] Phang, Y.C., Kassim, A.M. and Mangantig, E., (2021) Concerns of thalassemia patients, carriers, and their caregivers in malaysia: Text mining information shared on social media. Healthcare Informatics Research, 273, pp.200–213.

[12] Asmussen, C.B. and Møller, C., (2019) Smart literature review: a practical topic modelling approach to exploratory literature review. Journal of Big Data, [online] 61. Available at: https://doi.org/10.1186/s40537-019-0255-7.

[13] Gopal, L.S., Prabha, R., Pullarkatt, D. and Ramesh, M.V., (2020) Machine Learning based Classification of Online News Data for Disaster Management. 2020 IEEE Global Humanitarian Technology Conference, GHTC 2020

[14] Kaila, R.P. & Prasad, A.V.K., (2020) Informational Flow on Twitter - Corona Virus Outbreak – Topic. International Journal of Advanced Research in Engineering and Technology (IJARET), 113, pp.128–134.

[15] North, T.C. and Country, N., (2020) INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH EXPLORATORY DATA ANALYSIS AND TOPIC MODELLING ON TED TALKS Media Parichay Pothepalli. 22277, pp.54–59.

[16] He, T., Hong, W. and Hu, B., (2017) Evaluate the Chinese Version of Machine Translation Based on Perplexity Analysis. Proceedings - 2017 IEEE International Conference on Computational Science and Engineering and IEEE/IFIP International Conference on Embedded and Ubiquitous Computing, CSE and EUC 2017, 1, pp.278–281.

[17] Kherwa, P. and Bansal, P., (n.d.) Semantic N-Gram Topic Modeling. 726.

[18] Xu, G., Meng, Y., Chen, Z., Qiu, X., Wang, C. and Yao, H., (2019) Research on Topic Detection and Tracking for Online News Texts. IEEE Access, 7, pp.58407–58418.

[19] Sievert, C. and Shirley, K., (2015) LDAvis: A method for visualizing and interpreting topics. pp.63–70.Warmerdam, V., Kober, T. and Tatman, R., (2020) Going Beyond T-SNE: Exposing whatlies in Text Embeddings. pp.52–60.

[20] Cummins, J.A. and Nambudiri, V.E., (2022) Natural language processing: A window to understanding skincare trends. International Journal of Medical Informatics, [online] 160April 2021, p.104705. Available at: https://doi.org/10.1016/j.ijmedinf.2022.104705.