Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-03 | Issue-01 | March-2022

Automatic Topic Title Predicting from News Articles Using Semantic-Based NMF Model

Majid Khan[1], Imran Ali[2], Maheen Bakhtyar[3]
[1,2,3]Department of Computer Science & I.T University of Balochistan, Quetta, Pakistan.
majufone@gmail.com, imran.cs.uob@gmail.com, maheen2002@gmail.com

DOI: 10.5281/zenodo.6549853

ABSTRACT

Social medical being a predominant form of communication, millions of texts in terms of news articles, tweets, and snippets are generated worldwide every hour. From them discovering concise and useful knowledge has caught the interest from both academia and the business industry. Since the text document has an infinite amount of contextual data and it is sparse and ambiguous, therefore, learning topics automatically from them is a significant issue. To address this problem, this research paper proposes a semantic-based non-negative matrix factorization (NMF) model for extracting concise and meaningful topic titles for the text to grasp the whole text theme. The model is efficiently integrated with the semantic correlations between words and their context, which are learned through skip-gram. The NMF method is used to tackle this issue by using a block coordinate algorithm. In terms of topic coherence, extensive quantitative evaluations of the proposed models on a variety of real-world text datasets show that they outperform various state-of-the-art methods. The interpretability of these models demonstrated by qualitative semantic analysis, which identifies significant and consistent topics. It is an effective standard topic model for unstructured sparse text due to its superior performance and simple construction.

Keywords: Semantic-based NMF, Topic model, Semantic-correlation.

Cite as: Majid Khan, Imran Ali, Maheen Bakhtyar. (2022). Automatic Topic Title Predicting from News Articles Using Semantic-Based NMF Model. LC International Journal of STEM (ISSN: 2708-7123), 3(1), 75–87. https://doi.org/10.5281/zenodo.6549853

INTRODUCTION

Topic detection from a collection of documents has become a base for text mining applications such as text categorization, text summarization, topic modelling, and forecasting future trends. It can also be beneficial in many real-life applications, such as determining the importance of growing areas in the medical arena, as well as predicting their trend and popularity in the future. Similarly, one might discover popular trends and designs by looking at user reviews of a product. Similarly, mining financial tweets can used to forecast future stock market patterns. The underlying difficulties in all of these fields of text mining are how to develop significant topics and estimate future market trends.

Topic analysis also named (topic extraction, topic dictation, toping finding, or topic modelling) is a machine learning (data-mining) technique to extract meaning from text by recurrent topics or themes. It is used for organizing and understanding a large collection of unstructured text data **(Zhou, 2016)**, by assigning tags or categories to each document. Business deals every day with a huge quantity of unstructured text data in the form of emails, online reviews, etc. When it comes to analysing manually these huge amounts of text data, is nearly impossible to do, and it is time-consuming, tedious, and expensive. Manual sort of unsupervised data may likely to lead mistakes and inconsistency.

Background

Topic modelling **(Rehurek, 2010)** is a methodology for analysing the massive volume of data and extracting the hidden pattern, better decisions, optimizing internal processes, forecasting future trends, which makes it more efficient and productive. In addition, topic modelling is an ‘unsupervised’ machine learning technique **(Greene, 2008, Qiang, 2018, and Rehurek, 2010)** that does not require training, predefine list of labels, tags which is previously defined, it is a quick and easy way to analyse a large collection of data. However, there are many different approaches or techniques currently being used in text-mining for extracting the topics, with the most popular ones are Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA) **(Lee, 2009, Yeh, 2005, and Guan, 2018),** and Non-Negative Matrix Factorization (NMF). Unlike, LDA and LSA, NMF is one of the best, fast, and easily implementable for unsupervised and weakly supervised documents for predicting the topic.

Objective

In this paper, presents a complete framework for predicting the topic of a news article. Once the model is built, it will be possible to feed it a fresh text article and have it make predict the topic automatically. Before passing the new articles, the text data must be transformed using the TF-IDF and NMF models, and then the top predicted topic must be selected. The objective of this paper is to make topic modelling more accessible by presenting a practical approach in which a framework is offered and used on a case-by-case. In the next section, we will go through a literature review on topic modelling with non-negative matrix factorization.

LITERATURE REVIEW

Background Theory

As the mass amount of unstructured text data available on the web grows day by day, there is a pressing need to comprehend and extract valuable information from these massive amounts of data. Currently, there are two types of retrieval mechanisms available: the first is to explore documents based on subjects created by people, and the second is to retrieve documents based on a 'word query' such as sports topics, medical topics, and current news topics, among many others **(Han, 2000)**. Each browsing document requires a cluster label, which is subsequently used to classify the document and assign it to a browsing category. Semantic networks **(Berger, 2000)** and wordnet **(Carbonell, 1998)** are two examples of predefined topics for classifying articles. Earlier document retrieval research depended solely on keyword matching and vector-based representations, which are referred to as information retrieval.

As more and more unstructured text data became available, simple methods of information retrieval such as keyword extraction, vector representation, and clustering became inadequate, because these methods do not reveal high levels of semantics, such as the main theme of text documents.

As a result, it's necessary to find a topic from a collection of documents automatically. The names ‘topic prediction’ also ‘topic discovery’, ‘topic identification’, ‘topic finding’ all aim to find a basic theme of the document which carries semantic meaning **(Trang, 2017)**.

Previous Studies

Recent studies in the field of text mining for topic identifying from documents based on unsupervised approaches. There are few ones are most popular are latent semantics **(Erkan, 2004)**, and latent Dirichlet allocation **(Rehurek, 2010)**, graph-based methods **(Ordenes, 2014)**, MMR (maximal marginal relevance) [12], NMF model, and user feedback (UF) **(Yeh, 2005)**. Therefore, this research paper presents non-negative matrix factorization to redefine the patterns for the discovery of the topic. In the research paper of **(Wang, 2013 and Lee, 1999)** latent semantic analysis (LSA) is employed to extract features from document. A distributional semantics matrix is formed as a result of this approach, which connects sentences to the terms.

To relate the sentence and terms, a mathematical approach called Singular Value Decomposition is used. Their findings show that LSA for summarization of text provides a better result than the keyword-based method.

But one problem occurs while using SVD mathematical technique that it contains a positive and negative value which means it occurred some unimportant words may be put into the extracted summary published in the article **(Zha, 2002 and Röder, 2015)**, to solve this problem use nonnegative matrix factorization instead of latent semantic analysis.

METHODOLOGY

In this section, we will demonstrate our proposed semantic-based NMF method, where topics are learned from different texts corpus.

The proposed model uses word embedding to include semantic information in model training, allowing NMF to extract word co-occurrence from semantic links between words and their contexts (as shown in Figure.1)

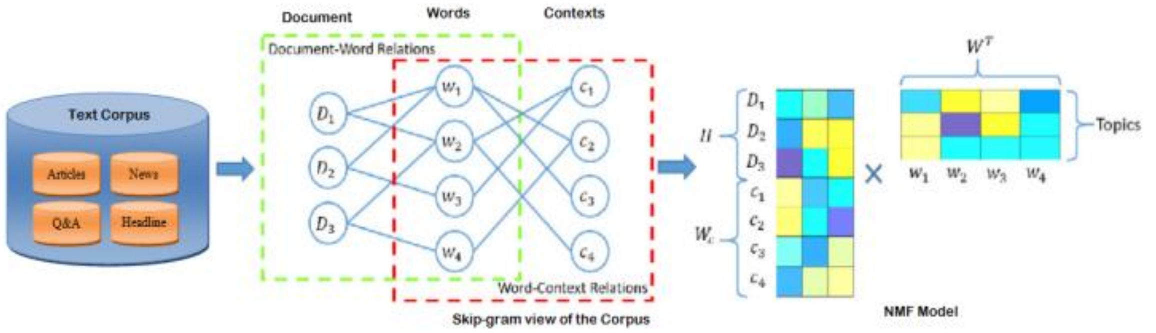

Figure 1: The suggested NMF model for topics learning from text dataset using a bi-relational matrix with connections between words and documents as well as words and context

Following are the major framework steps for prediction topic.

- Select the raw text data for predicting the model
- Apply NLP preprocessing functions like tokenization, stop word, lower case, POS, Stemming, and Lemmatization
- Applied non-negative matrix algorithm on preprocess text
- Extract the highest number of coherence score sentences for the model
- To train the model automatically on manual selected high score sentences
- Finally, predicting the topic of unseen documents

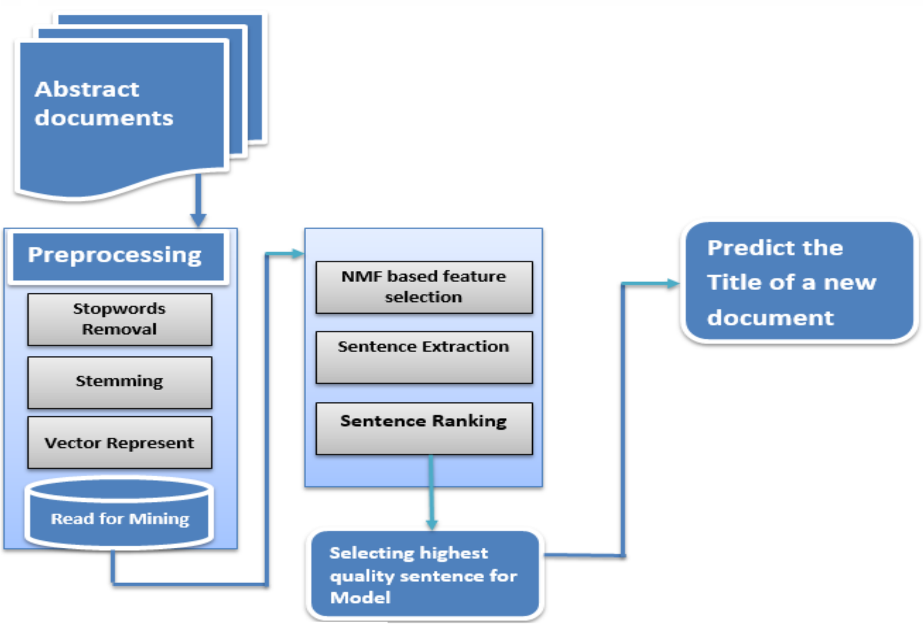

Figure 2: Flowchart of propose model

Data

This section will illustrate the performance of our models conducted comprehensive tests on a variety of real-world datasets which contain articles, news, headlines, tweets, and question & answers. To do this we implemented some state-of-the-art methods to compare the performance of our proposed on the same dataset to get better results.

CNN News: This data set was compiled from CNN news, which is full of text articles including titles and abstracts. These articles were taken from the CCN website from March to April 2020. After removing the stopwords, there are a total of 301 articles, the average word counted for this corpus is 732 and the standard deviation is 363. The articles on the web page cover a variety of topics such as investing, banking, success, video games, technology, markets, and so on.

Yahoo.Ans: The Yahoo! Answers Manner Questions is version of 2.04 dataset. The subjects of the Questions are collected from 10 different categories which include Finance, Fitness & Diet, and Tweets, etc. and so on.

20NG: This dataset is compiled from a subset of 20 News Group Corpus, in Term document format. It contains a variety of topics. The datasets have been preprocessed with stopwords, stemming, etc.

GoogleNews: This data was collected from word2vect website which contains English words around 3 million that have been embedded into a 300-dimensional latent space. Using the word2vec model on Google News corpus contain 3 billion running words.

Table 1: Datasets used in this paper have basic statistics

Dataset	#docs	#terms	density 'A'	density 'S'
CNN	301	9127	1.2861%	0.1369%
Yahoo.Ans	2225	2447	0.7693%	0.2677%
20NG	36392	2447	4.2667%	1.9494%
Tag.News	40754	4334	0.1997%	0.0973%

Model Development

Pre-processing the text data is the most crucial step before applying the non-negative matrix factorization method. When applying natural language processing techniques to an article's title to its corresponding abstract, it's important to ensure that the core theme of each sentence should be preserved, otherwise the final result will be different. Following are the few text processing functions which were used in this model:

- Sentence tokenization
- In lower cases the words
- POS part of speech
- Stemming (reduce to word stem)
- Punctuation, stop words, digits, single letters, and words with extra spaces should all be removed

NMF – Non-Negative Matrix Factorization

Also known as nonnegative matrix approximation, a state-of-the-art extraction method widely used for non-negative data feature extraction and dimensional reduction. The main difference between the nonnegative method and other factorization methods is non-negativity. It has been used in various fields of computer science for extracting hidden patterns and analyzing high-dimensional data into lower-dimensional spaces and effectively reducing the number of features, retaining the basic information to reconstruct the necessary original data. In other words, NMF is used for extracting meaningful semantic features automatically from a set of non-negative vectors. Paatero and Tapper first introduced it in 1994, then Lee and Seung's articles popularized it in 1999**.**

To extract latent structure from data, the NMF text normalisation technique is used. This technique is extensively used to reduce dimensionality by combining attributes to generate meaningful features.

NMF divides a data matrix A into two matrices W and H, each of which has no negative components. NMF is used to update W and H initial values in an iterative process until the product approaches A. The process ends when the number of iterations is reached or the error converges. In order to achieve the approximation of A is A ~ WH, the error function |A-WH| is minimised.

$$\frac{min}{W,H}\|V - WH\|_F^2 \qquad (1)$$

The research paper [9] explains how to calculate the 'W' and 'H' matrices. When using an NMF model, it maps the original data into a new set of features that the model discovers. NMF's inherent feature is that it automatically combines the data. As a result, it is appropriate for the summarization problem.

After applying the NLP preprocessing techniques of stop words remover, the document is presented as a set of sentences where term denote to each sentence of document. As seen in Figure 3 that matrix 'A' is decomposed into two smaller matrices 'W' and 'H'.

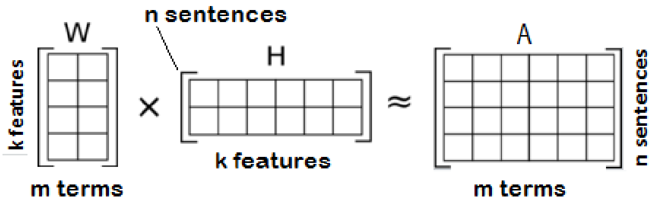

Figure 3: Text discovery through non-negative matrix factorization

A is the m x n matrix, W is the m x k matrix, and H is the k x n matrix, as shown in Figure 1. The number of articles to be created is denoted by m, the number of sentences in the article is denoted by n, and the number of features to be produced is denoted by k. In a sentence the term frequency is represented by each value in matrix A. The Generic Relevance of a Sentence - GRS is calculated using the 'H' matrix, as indicated in Equation (2). The sentences with the highest general relevance values are included in the summary.

$$\text{GRS of a j}^{\text{th}}\text{ sentence } = \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} \left(H_{ij}^{*}\ weight(H_{i*}) \right) \qquad (2)$$

Weight (H_{i*}) is calculated as below:

$$\text{Weight } (H_{i*}) = \frac{\sum_{q=1}^{n} H_{iq}}{\sum_{p=1}^{k} \sum_{q=1}^{n} H_{pq}} \qquad (3)$$

Feature Selection and More Reduction

After text processing, the next phase is to create features by converting text into vector form such as tf-idf. For feature selection, set the model's minimum and maximum word lengths 3 and 0.85, respectively, indicating that this model discards words with lengths less than 3 and greater than 85 percent in the documents. This will allow us to remove words from the model that aren't required. Besides tf-idf, vector we use skip-grams for weights that keep the range (1,2) which include unigram and bigram.

Evaluation of Topic

Assessing topic title from a collection of document is challenging due to their unsupervised learning process. For each corpus, there is no correct list of topics to compare the result with corresponding benchmark result or to calculate the error rate. This has sparked a surge of interest in the field of assessing the quality of topic models, and while much effort has been done to develop frameworks to address this issue, it remains an open research problem **(Röder M et.al, 2015)**.

The degree to which the learned topics match with the judgment of human is the key concern when using a topic model, as this is the goal for most use cases. Human evaluations of topic take long time, the concern is to find such measurements mechanism that are the most closely related to human judgment. One such measurement is topic coherence, which is discussed below.

Coherence Score

Using the coherence technique evaluates the relative distance of words inside a topic and assigns a score to a single topic because of degree of semantic similarity between the topic's high probability terms. Different sorts of coherence scores methods were employed in the paper of web search and data mining 2015 by **(Röder M et.al, 2015)**, but the two u_mass and c_v most prevalent methods. UMass is faster, used together with corpus and calculate the score with an intrinsic approach whereas c_v is found more accurate used together with the external corpus to calculate the score with an extrinsic matrix. In this paper, utilize the c-v approach to uncover cohesive topics.

$$Coherence\ (T) = \sum_{(t_i t_j) \in T} score(t_i t_j, \in) \qquad (4)$$

Automatically selecting the best number of suitable topics titles from an article using the coherence scores technique is not a very easy job while using the machine. To complete this task, the coherence technique is applied to a variety of topics, with the highest coherence score being chosen. Because the number of articles varies, this is simply the result of some trial and error. Each dataset composition is unique, needs a few manual commands to determine the topic range to search through. If the corpus has a large number of articles, running too many topics takes a lengthy time. However, the NMF model solves this problem and has an impact on the overall score of each topic. So finally, once the prediction model is built, it will be possible to feed it any type of text article and have it predict the topic. Before feeding the articles, they must be transformed using TF-IDF vectors and the NMF algorithm. The predicted models performed outclass on new text documents which were never seen by the model previously.

Method

After developing the proposed NMF semantic-based model, then we used some other different state-of-the-art methods for implementation and to get best performance and the purpose of generating evaluation report of our purpose model using these following methods for comparison:

Latent Dirichlet Allocation

LDA is one of the famous and well-known topic modelling technique which is being used for a large collection of raw text datasets. For comparison purposes, we use the implementation of LDA on different datasets.

Pseudo-document Topic Model

PTM is the short abbreviation of pseudo-document-based topic model, one of the recent new algorithms, which is being used for the aggregation of small texts without any additional information. It is also used for extracting topics from a short corpus

GPUDMM

The third method, which is used for comparison the result of proposed model, is GPUDMM. This method used semantic knowledge of external words vector from a huge corpus to promote semantically related words in each topic. For the implementation of this method, we use the Google News dataset as an external vocabulary dictionary. In experiments, the default number of topics is set K = 100. Values of $\alpha = 0.1$ and $\beta = 0.01$ for LDA. Used default hyper-parameter values for PTM and GPUDMM. PTM parameters for α , β and λ are set at 0.1 respectively and for GPUDMM is $\beta = 0.1$. For PTM, GPUDMM, and LDA two thousand iterations were run for Gibbs samples. For NMF kept the value of $\alpha = 1.0$ for CNN.News and Yahoo.Ans, respectively. To calculate density S, the value of κ and the value of γ are set to 1.0 to ensure consistency in the result.

DATA ANALYSIS AND RESULTS

Results

The topic coherence results of our models and different comparison approaches are shown in Tables 2. The bold font is used to highlight the top performance values, while the underline is used to indicate the second best performance values.

Table 2: Topic coherence results of different models using different datasets

	CNN.News	Yahoo.Ans	20NewsGroup	Tag.News
LDA	1.5048	**1.2957**	1.1637	**0.9346**
PTM	1.6628	1.1411	1.3745	0.8505
GPUDMM	0.9751	0.5798	0.9213	0.2815
NMF	**3.6318**	1.1394	**4.1477**	0.9184

Robustness Test

This section presents a comprehensive test of different on different text datasets, which validate the performance and confirm the results achieve by our proposed method over the other well-known methods which are used for comparison in this research study. The graphical results of propose method and other methods are given below:

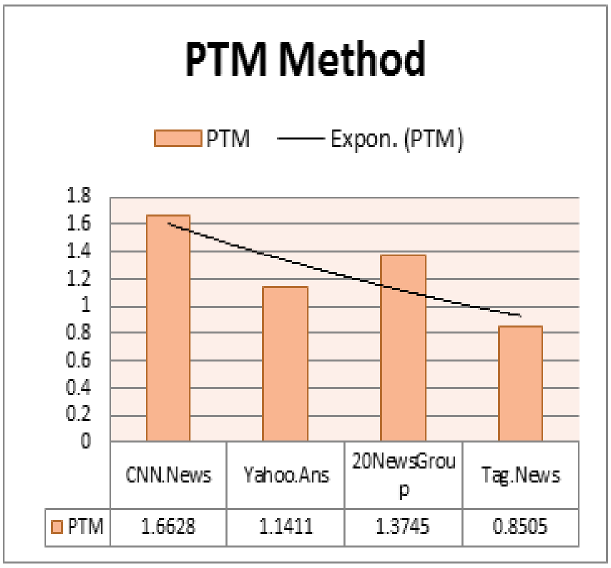

Figure 4(a): PTM results on different dataset

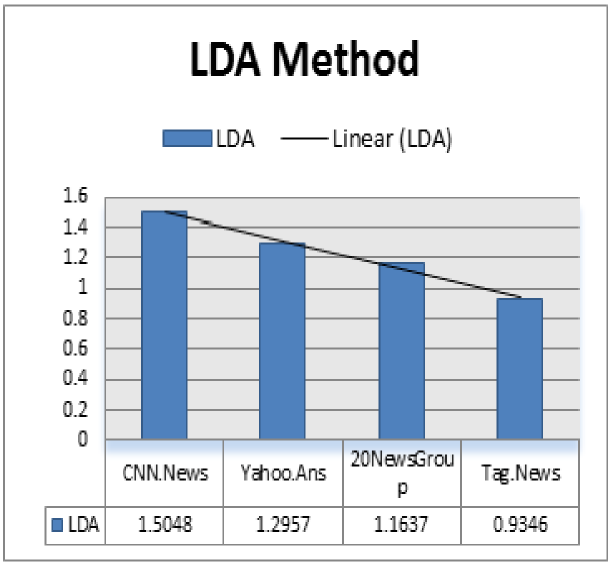

Figure 4(b): LDA results on different dataset

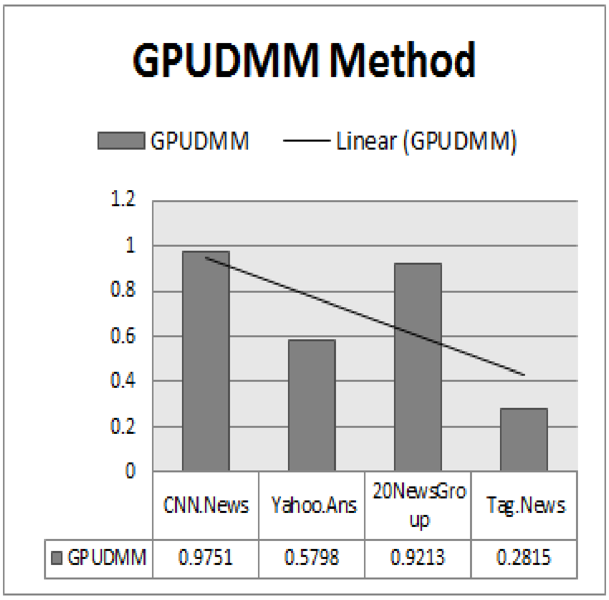

Figure 4(c): GPUDMM results on different dataset

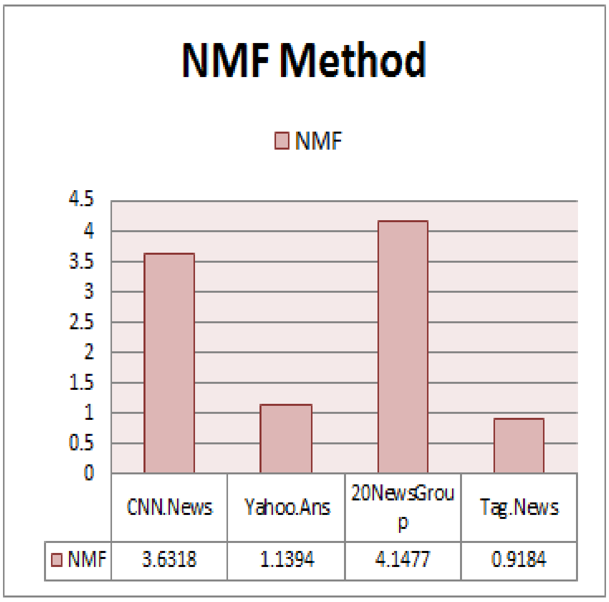

Figure 4(d): Propose NMF Model results on different real-world datasets

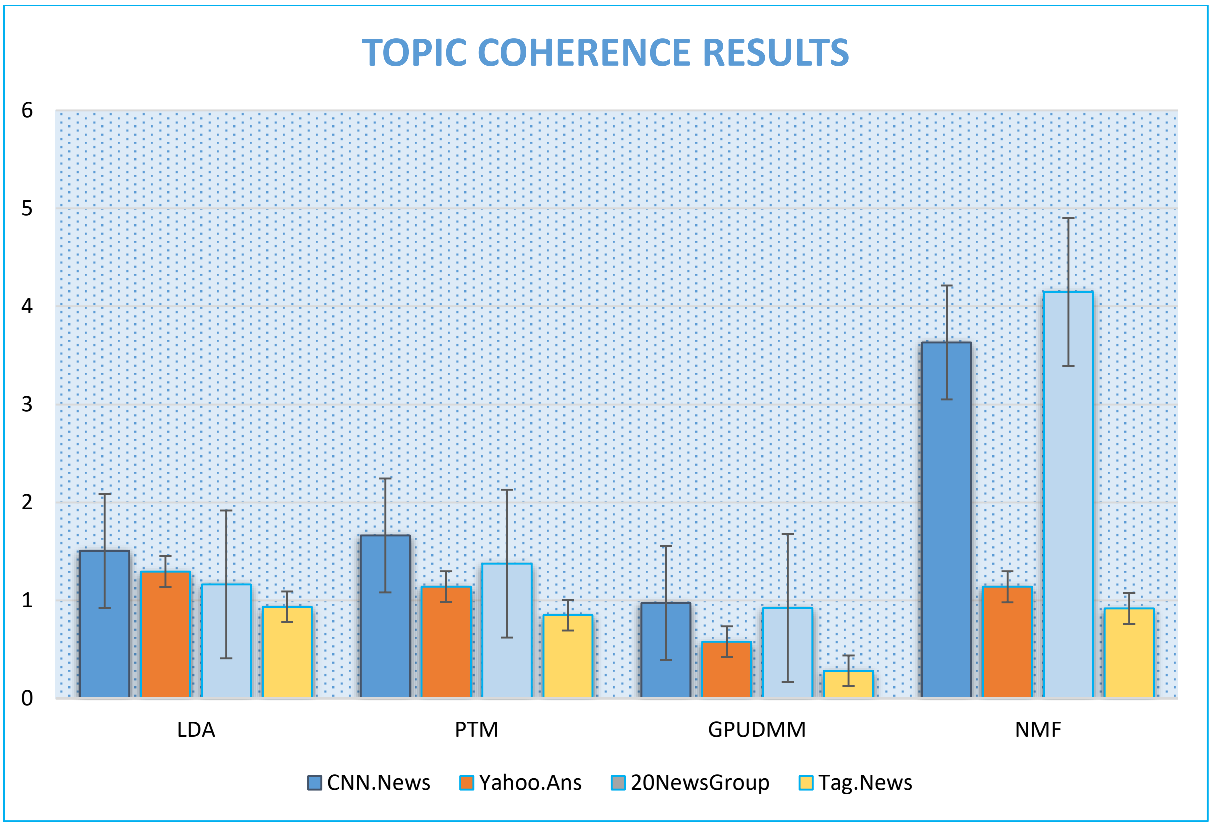

Figure 5: The performance of various state-of-the-art approaches was compared

Analysis

Our models outperform the standard NMF in Table (2), demonstrating that NMF is effective for learning concepts from texts. In comparison to LDA and current PTM, semantic-based NMF demonstrates big improvements, implying that our models find themes that are more cohesive. We visualize the top keywords in each topic for better understand GPUDMM's low performance in all circumstances. We discover that many top keywords are semantically associated, yet they do not tend to exist in the same article. Another explanation could be because the word semantic associations in Google News and other datasets differ, making Google News' general semantics knowledge ineffective for discovering subjects from these datasets.

CONCLUSION

This work proposed a unified way of working for automatically predicting the top of an article but obviously predicting the best number of suitable meaningful topics is critical, especially when it is discovered through the machine. After establishing the model different dataset documents were run on a machine and then picked the highest co-related topics. Another challenge was faced during the experiment phase were summarizing all the machine selected topics, the best solution was found here to have human go through but this is not ideal. Another way is found, to use the words in each topic that has the highest score for that topic and map those back to the features Overall it does a good job of predicting topics while using different datasets. In the future, the next step will be a more detailed inspection to make this process purely machine-based rather than human intervention in any way for predicting good topics.

REFERENCES

[1] Zhou, D., Xu, H., Dai, X. Y., & He, Y. (2016, July). Unsupervised Storyline Extraction from News Articles. In IJCAI (pp. 3014-3021).
[2] Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In In Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks.
[3] Greene, D., Cagney, G., Krogan, N., & Cunningham, P. (2008). Ensemble non-negative matrix factorization methods for clustering protein–protein interactions. Bioinformatics, 24(15), 1722-1728.
[4] Qiang, J., Li, Y., Yuan, Y., & Liu, W. (2018). Snapshot ensembles of non-negative matrix factorization for stability of topic modeling. Applied Intelligence, 48(11), 3963-3975.
[5] Lee, J. H., Park, S., Ahn, C. M., & Kim, D. (2009). Automatic generic document summarization based on non-negative matrix factorization. Information Processing & Management, 45(1), 20-34.
[6] Yeh, J. Y., Ke, H. R., Yang, W. P., & Meng, I. H. (2005). Text summarization using a trainable summarizer and latent semantic analysis. Information processing & management, 41(1), 75-95.
[7] Guan, J., Levitan, A. S., & Goyal, S. (2018). Text mining using latent semantic analysis: An illustration through examination of 30 years of research at JIS. Journal of Information Systems, 32(1), 67-86.
[8] Han, K. S., Baek, D. H., & Rim, H. C. (2000, November). Automatic text summarization based on relevance feedback with query splitting. In Proceedings of the fifth international workshop on Information retrieval with Asian languages (pp. 201-202).
[9] Berger, A., & Mittal, V. O. (2000, October). Query-relevant summarization using FAQs. In Proceedings of the 38th annual meeting of the association for computational linguistics (pp. 294-301).
[10] Carbonell, J., & Goldstein, J. (1998, August). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 335-336).
[11] Trang, N. T. T., Huong, L. T., & Hung, D. V. (2017, December). Enhancing extractive summarization using non-negative matrix factorization with semantic aspects and sentence features. In Proceedings of the Eighth International Symposium on Information and Communication Technology (pp. 78-83).
[12] Erkan, G., & Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of artificial intelligence research, 22, 457-479.

[13] Ordenes, F. V., Theodoulidis, B., Burton, J., Gruber, T., & Zaki, M. (2014). Analyzing customer experience feedback using text mining: A linguistics-based approach. Journal of Service Research, 17(3), 278-295.
[14] Wang, Y., & Ma, J. (2013, November). A comprehensive method for text summarization based on latent semantic analysis. In CCF International Conference on Natural Language Processing and Chinese Computing (pp. 394-401). Springer, Berlin, Heidelberg.
[15] Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), 788-791.
[16] Zha, H. (2002, August). Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 113-120).
[17] Röder, M., Both, A., & Hinneburg, A. (2015, February). Exploring the space of topic coherence measures. In Proceedings of the eighth ACM international conference on Web search and data mining (pp. 399-408).

BIOGRAPHY

Majid Khan received the B.cs degree in Computer Science from Government Science College, Balochistan, Quetta Pakistan in 2006. He completed his M.cs degree from department of Computer Science University of Balochistan, Quetta Pakistan in 2009. He is currently working as Research Assistant in UoB, Quetta, Pakistan. His current research area include Machine Learning, Text Mining, Automatic Text summarization, and Topic Modelling.

Imran Ali serving as lecturer in Department of Computer Science and Information Technology since 2007. He completed MS in Computer Science from Asian Institute of Technology Bangkok, Thailand in 2010. His research interest include Automatic Text Summarization, Frequent Pattern Mining, and Named Entity Recognition.