Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 40 Issue: 09-2021
DOI 10.17605/OSF.IO/ F6JVP

AN EFFICIENT NOVEL STRATEGY FOR ONLINE SOCIAL NETWORKS OF A Q&A COMMUNITY FORUMS USING TOPIC MODELLING METHODS

P. Venkateswara Rao[1] and A.P. Siva Kumar[2]

[1]Research Scholar of Computer Science and Engineering, Jawaharlal Nehru Technological University Ananthapur, Anantapuramu-515002, Andhra Pradesh, India. Mail Id: pvenkat2004@gmail.com

[2]Associate Professor of Computer Science and Engineering, Jawaharlal Nehru Technological University Ananthapur, Anantapuramu-515002, Andhra Pradesh, India. Mail Id: sivakumar.ap@gmail.com

Abstract: The utilization of customer-produced data collected by society broadcasting to scrutinize public estimation and systematic interaction on hire and protection issues is an emerging trend in technical study. This analysis of the data, as well as the introduction of a social question-and-answer website, is part of a larger package aimed at determining the elements that impact community preferences for technological knowledge and thoughts. This study measured the effect of the response stylistic and supporting functions on the number of appointments received with the response using a web search engine, subject modelling, and degeneration data modelling. The results of the model reveal that Quora users are more inclined to talk just about technology when compared to earlier studies based on open evaluations. It may possibly fail if the query's keywords do not match the text content of huge texts including pertinent queries on or after previous techniques, such as CNNMF and NMF, as well as some constraints. Furthermore, consumers are frequently inexperienced and offer vague requests, resulting in mixed outcomes and issues with existing approaches. To address this issue, we offer a Hadoop model for finding topics for short texts that is distributed using semantics and non-negative matrix factorization (HDiSANNMF). It efficiently combines the semantic associations of the word context obsessed by the model, which studies the semantic relationships between words and their context without ignoring the corpus's grammatical shape. The researchers are attempting to rearrange the major findings and propose new ways for modelling distributed themes in order to deal with increasingly complex technologies and platforms, as well as the amount of time and space required to build the model. This portion gives a short-term overview of the structure of public queries and replies from around the world, as well as real-time tracking of the primary issues of accommodation and work opportunities for next-generation technology.

Keywords: LDA, NMF,Hadoop, Topic models, quora, stack overflow, Twitter-API, NLTK.

1. Introduction

Stack Overflow has grown in popularity as a software development resource over the last decade. Inexperienced programmers now turn to Stack Overflow for answers to their questions on software development. Question labels are used by sites like Stack Overflow and Code Review to match them with persons who can answer them. It's possible that new Stack Overflow users or inexperienced engineers will not accurately flag their posts. Even if the issue is interesting and

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 40 Issue: 09-2021
DOI 10.17605/OSF.IO/ F6JVP

adds value to the community, this results in moderators voting and tagging postings. [25-26].

Every day, a large volume of brief text is produced, including tweets, examination queries, questions, image tags, keywords, titles, and so on. They have had a significant impact on our daily lives. Discovering information on the issue becomes an exciting but difficult research activity that draws a lot of interest. As a result, research and development of large data processing frameworks is fast rising [16]. One of the promising open-source software frameworks [17] is the subject of this comment. Hadoop is a distributed computing platform for executing huge everyday computations in graphs, matrices, deep learning, machine learning, and network algorithms, based on group synchronous equivalent operations. It's written in Java and runs on the Hadoop Distributed File System (HDFS), thus it works with Hadoop clusters seamlessly. Websites like Stack Overflow are frequently used by software engineers and programmers to find answers to their queries. This data [1] can be used to determine which components of programming and APIs are the most difficult to grasp. To classify redundant stack problems that are of relevance to two overlapping views, which are programming principles and the type of information sought, in this comment. The flood assault contains a significant amount of data on a variety of computer programming issues.

2. Related Works:

The Info Lab group at MIT's Computer Science and Artificial Intelligence Laboratory developed a Start method for open question and answer software systems. However, question-and-answer systems that are utilized to complete a course's assignments are extremely unusual. As a result, an intelligent question-and-answer system that returns replies to consumer questions based on exchange rate principles [3] [19] has been built. We show how open vocabulary fits into natural language processing and how, when combined with machine learning, it can be used to infer a person's personality from their use of language during a job interview. Using data from over 46,000 participants who participated in an open interview and answered a question [20].

In connection to the subjects covered, the usage of online reviews may be biased. When it comes to the types of customers who leave reviews, we expect prejudice, especially when it comes to extreme ones. Customers who have had an exceptionally fantastic or extremely negative experience, for example, are more inclined to leave an online review to recommend or warn others about the facility. Although this would likely aid in distinguishing between themes of interest, it is possible that this represents a deviation in knowledge of the challenges given in terms of a proclivity towards extremes [21] [22]. However, this does not give you a solid sense of how to model themes. Furthermore, we are unable to receive SymNMF document submissions directly. As a result, the method suggested in this paper is the first to examine the building of a standard model for short texts based on MFN [23] [24].

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 40 Issue: 09-2021
DOI 10.17605/OSF.IO/ F6JVP

3. Proposed Architecture:

NLP (Natural Language Processing) to the acquisition of immeasurable information in technical domains that are strongly driven to examine the models are all examples of thematic models. The unreliable number of TM subjects associated with bad performance in the placement assistant technical field is facilitated by TM topic analysis, which is a big problem. In this way, the required visuals are an important element of clipping data in order to define the cluster's direction for the assistant's task. As a result, to believe and contribute to the anticipated Hadoop thematic representations, which are disseminated with Hadoop non-negative matrix factorization (NNMF) and Hadoop distribution with latent Dirichlet distribution (DLDA), the correct approaches to balance and align to the Addressing the questions and answers to pairs or topics or terms or from different data sources in the technical data set to be placed. These proposed models can be found in the table below.

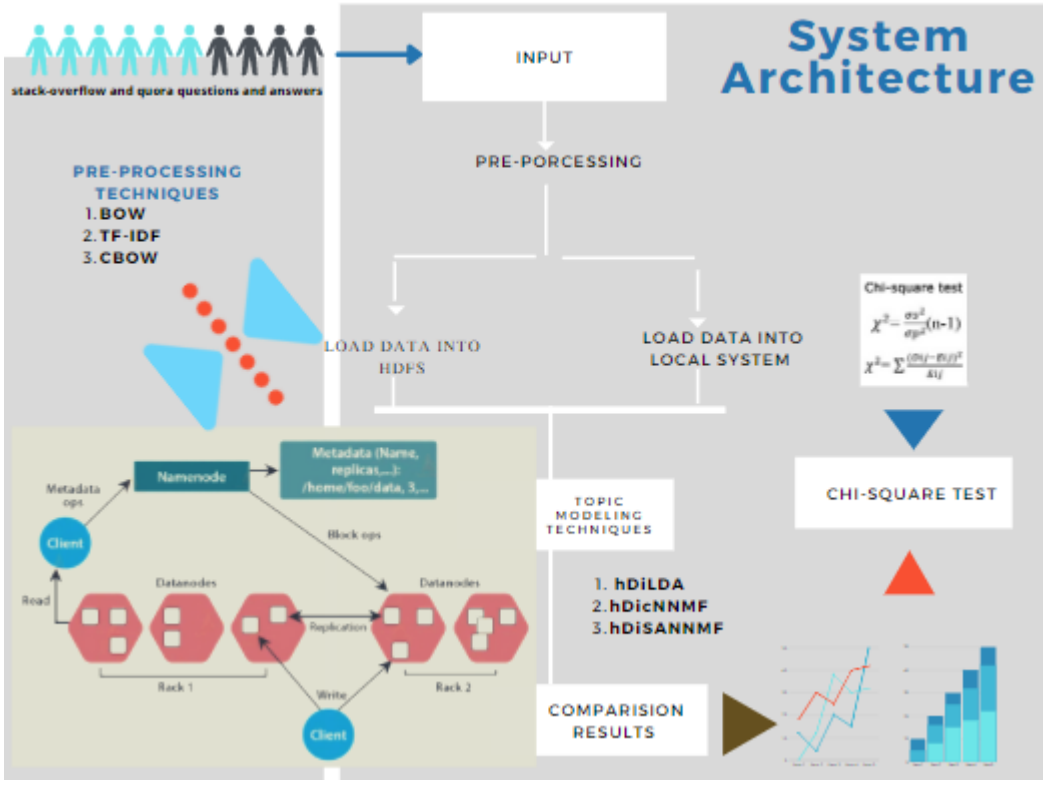


Fig 1: Proposed Architecture

For each observation in the data set, the roles are different as unique and

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 40 Issue: 09-2021
DOI 10.17605/OSF.IO/ F6JVP

measurable traits / properties. These qualities are the context / category / marker that determines the ability of the query text accurately and uniquely. The identical function can be used to the probable content of the inquiry, despite being introduced in the machine learning algorithm, searching for learning patterns with intent. These functions, on the other hand, essential be vectorized consequently that the algorithm can enhance and reduce losses when defining models [2].

Thematic prototypes are geometric probability models that use simple de-composition (SVD) and latent Dirichlet mapping to identify obscure semantic structures in a text (LDA). The primary property by which we conclude the purpose extraction is the body of the query. This framework covers Hadoop distributed theme modelling methodologies for the best results.

3 .1 Methodologies

In the issue of stack overflow, our investigation responded to user confusion by first assigning the question the suitable identifier or intent and then matching the user question mark with the most relevant question. The answers to the first ten most important questions were finally available. We'll cover some background material in this section, as well as the block coordinate descent method and how it's employed in NMF for topic modelling. Then, using our Sea-NMF model and a block-coordinate descent strategy, we will guesstimate latent representations of phrases and fleeting texts.

3.2 Data Collection

The data comes from Stack Overflow's official blog view and public atmosphere assessment. The control includes roughly 2.98,000 user records. Stack Overflow runs a poll every year to establish its users' interests, and the results can be utilized for various analyses. I begin by attempting to obtain data from Stack Overflow, beginning with the logs. When compressed, however, the collection file is 12 GB in size.

Due to the restrictions of our restricted processing hardware and our incapacity to work with such a large dataset, we chose to use the Kaggle dataset in this situation. In Kaggle, Python projects have their own data collection. Hadoop is a free open-source program that lets you process enormous volumes of data using the HDFS file system. Map reduction refers to the managing and training paradigms for Java, Python, R, and Spark-based computational activities. In the Map Reduce approach, both Map and Map Reduce are important fields. The target divides individual items into tuples (keys / key values), which accepts one set of data and changes it to another. [19].

3.3 Preprocessing and Data Preparation:

The Word embedding is a language modelling technique in which words are converted into real-number vectors. It's a multi-dimensional vector space that represents words or sentences. Word embeddings can be created using a variety of techniques, including neural networks, co-occurrence matrices, and probabilistic

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 40 Issue: 09-2021
DOI 10.17605/OSF.IO/ F6JVP

patterns. Word2Vec is a set of word embedding models that use one-dimensional two-layer neural networks with one input layer, one hidden layer, and one output layer. The CBOW model predicts the current word based on context words in a specific timeframe. The context words are in the input layer, while the current word is in the output layer.

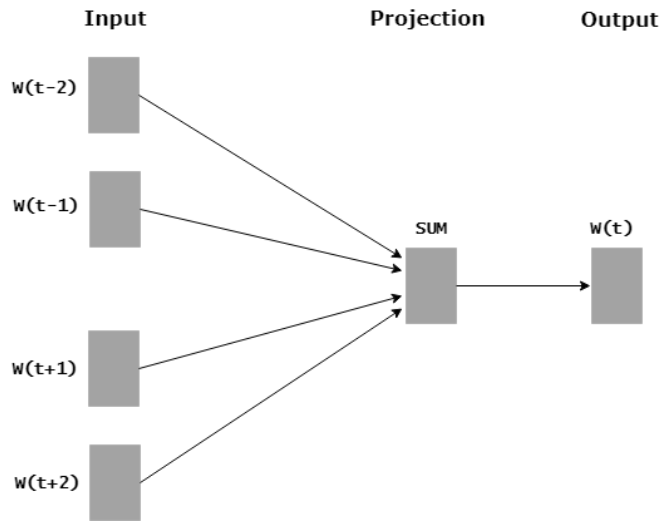
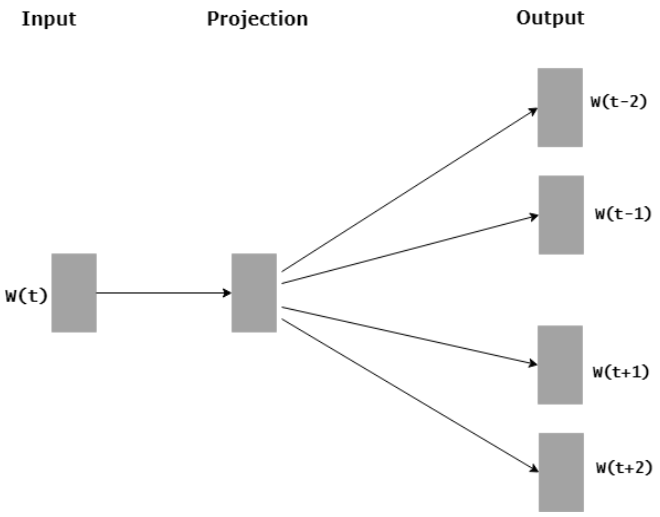


Fig 2: CBOW Model -1

The Skip Gram estimates the adjacent context words surrounded by a set window of opportunity given a current word. The input layer encompasses the most recent word, whereas the output layer encompasses the context words. The hidden layer stores the number of components in which want to correspond to the modern word at the input layer.



Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 40 Issue: 09-2021
DOI 10.17605/OSF.IO/ F6JVP

Fig 3: Skip Gram model

The underlying concept behind word embedding is that words that appear in comparable contexts tend to be stronger in vector space. The modules nltk and genism are required to generate word vectors in Python.

If you're looking for tf-idf, you're presumably already familiar with characteristic extraction and what it is. In terms of word retrieval, it's one of the most essential strategies for describing how relevant a word or phrase is to a document.

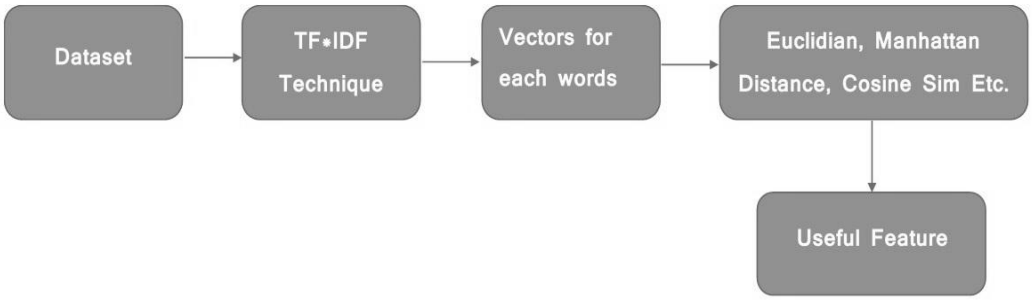


Fig 4: working flow of TF_IDF technique

Tf*Idf does not turn unprocessed data into usable features directly. To begin, it transforms raw strings or datasets into vectors, with each word having its vector. Then, for obtaining the feature, we'll utilise a technique like Cosine Similarity, which works on vectors, and so on. We can't just feed the string to our model, as we know. As a result, tf*idf supplies us with numeric values for the entire document.

3.4 Topic modeling:

The Topic modelling is a kind of abstract patterning that is used to find abstract "themes" in document collections. The concept is that we will do unsupervised categorization on various papers, resulting in the discovery of some natural topic groups. Using topic modelling, we can answer the following question.

1. What is the document's principal idea or topic?

2. Can we find another document with a comparable topic if we have one?

3. How has the field of themes changed over time?

Topic modelling can aid in the search process optimization. We'll talk about Latent Dirichlet Allocation, a topic modelling approach, in this article.

3.4.1 Distributed Latent Dirichlet Allocation (DLDA)- The latent Dirichlet distribution or allocation is a difficult Bayesian technique, and the latent word refers to capturing the meaning of the text in order to uncover hidden terms or themes of the words in the corpus of each document in the corpus [11]. Latent Dirichlet allocation is one of the most used approaches to topic modelling. Each document has a few terms, and each topic is associated with a set of words. The LDA's

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 40 Issue: 09-2021
DOI 10.17605/OSF.IO/ F6JVP

purpose is to use the words in the document to determine which subjects it belongs to. It is anticipated that texts with similar topics will use the same terminology. This allows the documents to map the latent topic probability distribution to the topic probability distribution.

Setting up Generative Model:

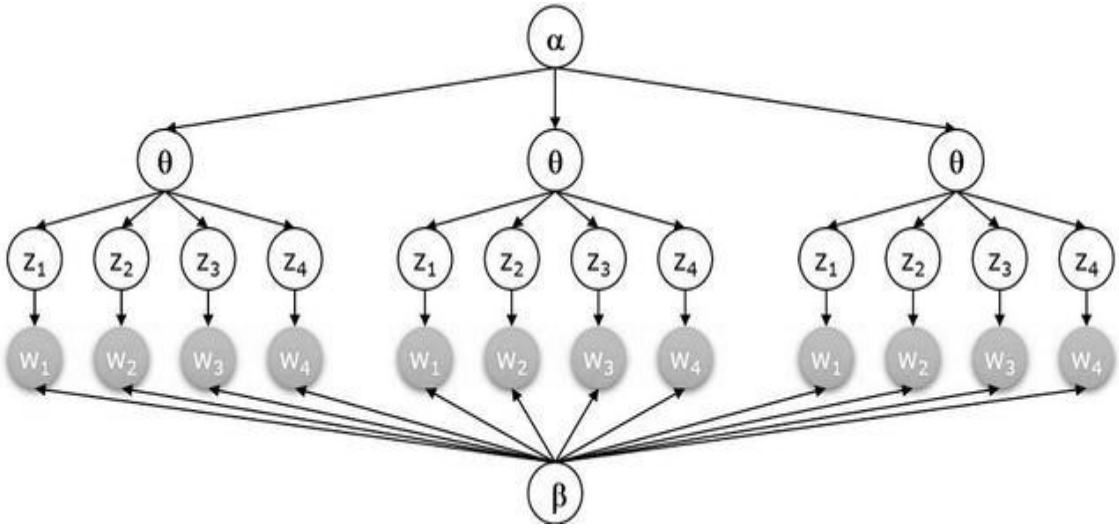


Fig 5: Generative Model

• Assume we have D documents that use the V-word kinds' vocabulary. Each document is made up of N tokens (can be removed or padded). Now that we've assumed K topics, we'll need a K-dimensional vector to reflect the document's topic distribution.

• With a common symmetric prior, each topic has a V-dimensional multinomial beta k over words.

3.4.2 NNMF:

The Non-Negative Matrix Factorization For a matrix A with dimensions m x n and each element equal to 0, NMF can factorize it into two matrices W and H, each with dimensions m x k and k x n and only non-negative components. Matrix A is defined as follows:

$$A_{mXn} = W_{mXk} H_{kXn}$$

where,

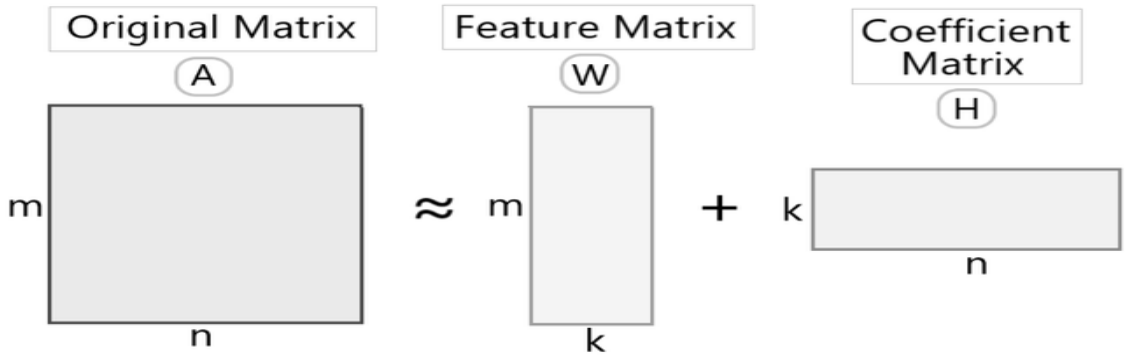
A -> Original Input Matrix **(Linear combination of W & H)**

W -> Feature Matrix

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 40 Issue: 09-2021
DOI 10.17605/OSF.IO/ F6JVP

H -> Coefficient Matrix **(Weights associated with W)**

k -> Low rank approximation of A **(*k* ≤ *min(m,n)*)**



The goal of NMF is to reduce dimensionality and extract features. When the lower dimension is set to k, the purpose of NMF is to find two matrices with solely nonnegative entries, W Rmk and H Rnk.

The NMF (Nonnegative Matrix Factorization) is an unsupervised learning technique that requires been effectively employed in a variety of domains, including signal processing, face recognition, and text mining. NMF's capacity to extract useful information from high-dimensional data.

The distributed visual NMF is a collection of linear algebra algorithms for extracting the latent structure in data intended as a non-negative matrix, and it is used for topic models in the Hadoop environment for distributed the tweet data corpus [9][10]. First, non-negative matrix factorization can be shown to be equal to optimizing the identical objective function as probabilistic latent semantic analysis. Using scikit-learn, the following algorithm NMF [12][13] addresses topic modelling on a term-document matrix.

3.4.3 SANMF:

We offer a new model for modelling subjects with short text that is supported by NMF semantics (HdiSANMF), as shown in fig. 3. The texts, words, and context are labelled DI, WI, and CI, respectively, in this diagram. Our target job combines the advantages of both the NMF topic modelling model and the grammar skip capture model in the context of word semantic correlations, and the proposed HDiSANNMF model can capture the semantics of short text bodies based on word-document and word-context associations The vector images of documents, situations, and words in latent space are shown in Figures H, Wc, and W. Each W column symbolizes a different theme.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 40 Issue: 09-2021
DOI 10.17605/OSF.IO/ F6JVP

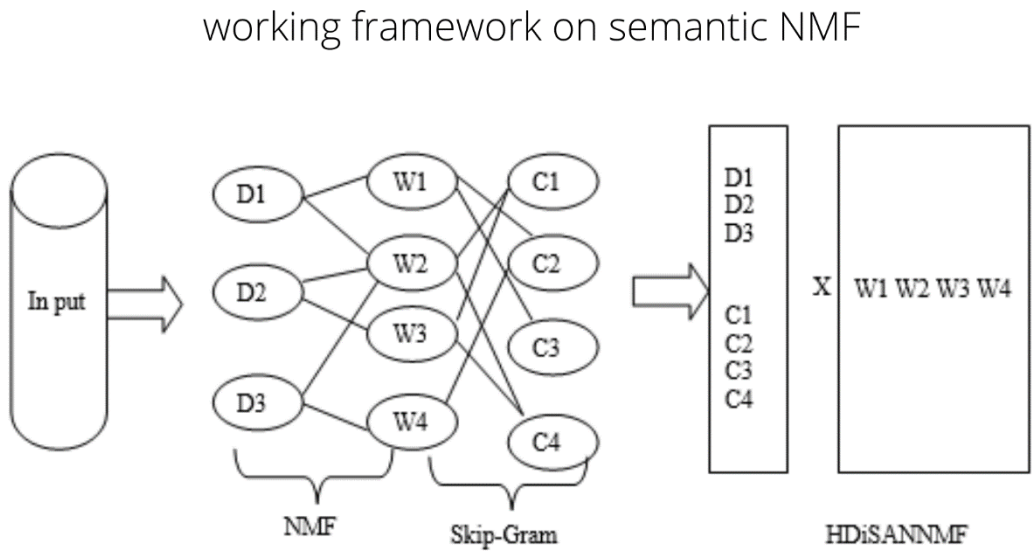


Fig3: working framework on Sea-NMF

The document term matrix "A" is first built using a representation of a "bag of words" as part of the HDiSANMF algorithm. The semantic correlation matrix S is then calculated, using the latent factor matrices W, Wc, and H arbitrarily adjusted with non-negative real integers. Then, surrounded by for each reiteration, columns will update your coordinates.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 40 Issue: 09-2021
DOI 10.17605/OSF.IO/ F6JVP

Algorithm:HDiSANNMF

Step:1 **Input**

a) TDM:Term Document Matrix A
b) Semantic co-relation matrix S
c) No of Topics K,alpha(α)

Step:2 **Output:**

W, W_c, H

Step:3 **Initialize:**

$(W, W_c, H)>=0$(Zero) and t=1

Step:4 **Repeat the process**

For K=1,K do

a. Calculate W^t
b. Calculate W^t_c
c. Calculate H^t

Exist

Until Converge

3.4.4 Experimental Results and Discussion:

The purpose of this research was to assess a few resemblances factoring models that were used to locate comparable questions make use of topic modelling methods, and then to generate the most relevant questions that were effectively derived. To obtain the 10 most relevant questions for each user query, make use of cosine similarity for k topics, topic modelling, and ensemble models to quantify the importance of the similarity of the questions via a call and use cosine similarity for k topics, topic modelling, and ensemble models. The precision, recall, accuracy, and F-1 Score of the Topic Modeling technique applied for similarity metrics are shown in Figures 5.and It enables Hadoop Distributed Semantic Assisted Non-Negative Matrix Factorization, which improves the model's accuracy from real-time corpus data as compared to other topic models. Here are the precision and recall comparative metrics produced by TP,TN,FP, and FN in topic modelling models.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 40 Issue: 09-2021
DOI 10.17605/OSF.IO/ F6JVP

Table 1: Hadoop Distributed semantics assisted NNMF

Q&A Corpus	Precision (P)	Recall ®	Accuracy	F-1 Score
1	1.00	1.00	1.00	1.00
2	0.98	0.98	0.99	0.98
3	0.96	0.96	0.97	0.96
4	0.94	0.94	0.96	0.94
5	0.91	0.91	0.95	0.91
6	0.89	0.89	0.93	0.89
7	0.86	0.86	0.92	0.86
8	0.84	0.84	0.91	0.84
9	0.81	0.81	0.89	0.81
10	0.78	0.78	0.88	0.78

Fig 4: precision, Recall, Accuracy and F-1 Score of Distributed hadoop SANNMF

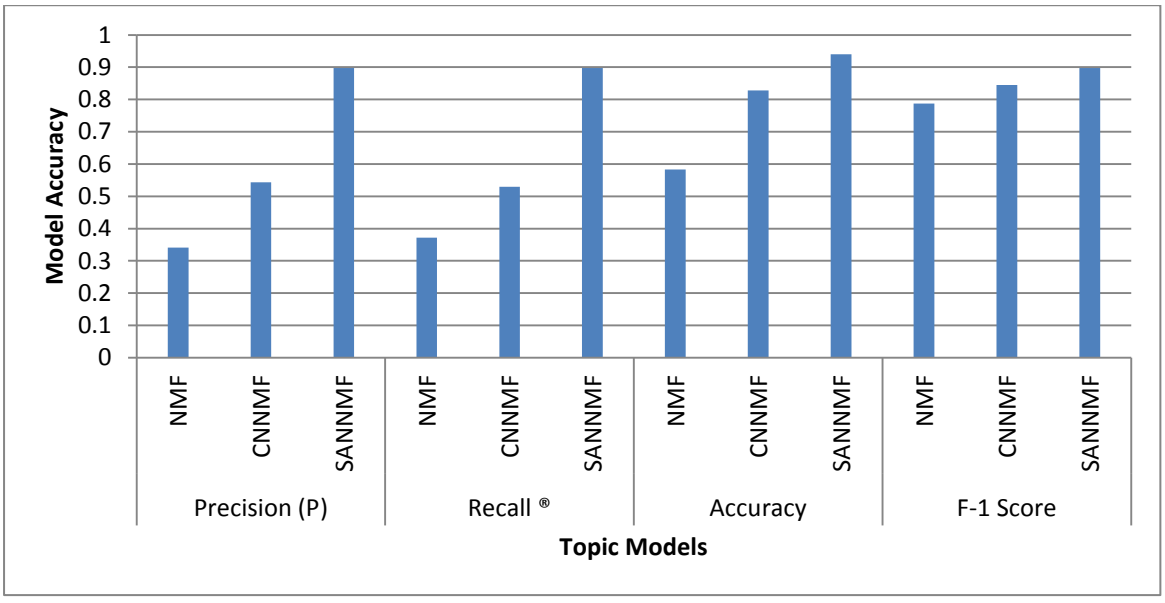


Fig: 5 Comparison results on Topic modeling techniques

The precision, recall, accuracy, and F-1 Score of the Topic Modeling technique applied for similarity metrics are shown in Fig 4 and, and it enables the Hadoop

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 40 Issue: 09-2021
DOI 10.17605/OSF.IO/ F6JVP

Distributed Semantic Assisted Non-Negative Matrix Factorization for better model accuracy from real-time corpus data when compared to other topic models. Here are the precision and recall comparative metrics produced by TP, TN,FP, and FN in topic modelling models.

Topic models like as the NMF, CNNMF, and HDiSANNMF are used to analyse data corpora and are balanced in Topic Modeling Techniques. As shown in Figure 5, Hadoop Distributed Semantic Aided Non-Negative Matrix Factorization balanced the comparison analysis. Finally, this book provides a brief overview of public question-and-answer structures around the world, as well as a timeline of the major subjects. Real-time scrolling of housing and work options for next-generation technology around the world. In table 11, the Chi-square test is a non-parametric method for comparing the association between two category or nominal issues. For example, if we have distinct words and term outcomes (cured and noncured), we may apply the chi-square test for independence to see if themes or terms are associated to community question and answer outcomes.

4 Conclusion & Future work:

As part of a modern study with statistics set and techniques, researchers look at how topic Models (TM) convey information about technical terminology in the context of professional concerns. Stack Overflow is becoming increasingly relevant in the area of sociological research and communication advisers. This may result in the loss of vital information. When modelling a subject or words using the HDiNNMF method with precision, recovery, and F-1 indications compared to previous results, combine cluster modelling with topic modelling to build an overall presentation and receive superior results. Believe that the presented technique, which incorporates valuable question-and-answer data as well as a quantitative analytical and qualitative outcomes procedure, demonstrates that HDiSNNNMF's best keywords lead to more semantically associated subjects. As a result, we believe that the presented strategy is a good topic model for this project.

In the future, scientific statements and community feedback will be used to facilitate task linking with RPA PEGA Robotics and the AWS Eco-system to automate manual work with streaming questions and responses. This study briefly illustrates the community composition of queries and responses throughout the world, as well as the real-time evolution of the main subjects of hosting opportunities and utilization of next-generation technology.

References

1. D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” the Journal of machine Learning research, vol. 3, pp. 993–1022, 2003.
2. LI Hua-Meng,LI Hai-Rui,XUE Liang. TFIDF Algorithm Based on Information Gain and Informati Entropy[J]. Computer Engineering, 2012, 38(08): 37-40.
3. Hanchen Jiang, Maoshan Qiang, Dongcheng Zhang, Qi Wen, Bingqing Xia, Nan An. "Climate Change Communication in an Online Q&A Community: A Case Study of Quora", Sustainability,

ISSN：1671-5497
E-Publication：Online Open Access
Vol: 40 Issue: 09-2021
DOI 10.17605/OSF.IO/ F6JVP

2018
4. Campbell, J.C., Hindle, A. and Stroulia, E., 2014. Latent Dirichlet allocation: extracting topics from software engineering data. In The art and science of analyzing software data (pp. 139-159). Morgan Kaufmann
5. Rainer Lienhart, Stefan Romberg, and Eva H¨orster. Multilayer pLSA for multimodal image retrieval. In Proceeding of the ACM International Conference on Image and Video Retrieval, CIVR '09, pages 9:1–9:8, New York, NY, USA, 2009. ACM.
6. S. Arora, R. Ge, R. Kannan, and A. Moitra. Computing a nonnegative matrix factorization provably. In Proc. the 44th Symposium on Theory of Computing (STOC), pages 145–162,2012.
7. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Annual Conference on Neural Information Processing Systems, pp. 556–562 (2000)
8. Yan X, Guo J Learning topics in short text using ncut-weighted non-negative matrix factorization on term correlation matrix,2013
9. Huang L, Ma J, Chen C (2017) Topic detection from microblogs using T-LDA and perplexity. In: 24th Asia-Pacific software engi- neering conference workshops,2018
10. W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In Proc. the 26th Annual International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR), pages 267–273, 2003
11. Peng Zhang, Department of Mathematics, Zhejiang University, Hangzhou, 310027 China; Wanhua Su Statistical inference on recall, precision and average precision under random selection,2012 7,Print on Demand(PoD) ISBN: 978-1-7281-4715-4,2019
12. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. BMC Bioinformatics 2010; 11:367.
13. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999; 286:531–7.
14. Dai L, Gao X, Guo Y, Xiao J, Zhang Z. Bioinformatics clouds for big data manipulation. Biol Direct 2012; 7:43.
15. V. K. Vavilapalli, A. C. Murthy, C. Douglas, S. Agarwal, M. Konar, R. Evans, T. Graves, J. Lowe, H. Shah, S. Seth, B. Saha, C. Curino, O. O'Malley, S. Radia, B. Reed, and E. Baldeschwieler, "Apache Hadoop YARN: Yet Another Resource Negotiator," In Proc. ACM
16. F. H. Gebara, H. P. Hofstee and K. J. Nowka, "Second-Generation Big Data Systems," IEEE Computer, vol. 48, no. 1, pp. 36-41, 2015.
17. Apache Hama, accessed on June 16, 2016. [Online]. Available: https://hama.apache.org/
18. Yeung, K. Quora Now Has 100 million Monthly Visitors up from 80 million in January. Available online: http://venturebeat.com/2016/03/17/quora-now-has-100-million-monthly-visitors-up-from-80-million-in-january/ (accessed on 28 March 2016).
19. Alexa Webpage: Web Traffic Statistics of Quora. Available online: http://www.alexa.com/siteinfo/www. quora.com (accessed on 15 April 2016).
20. M. Jayaratne, B. Jayatilleke: Predicting Personality Using Answers to Open-Ended Interview Questions, Digital Object Identifier 10.1109/ACCESS.2020.3004002. July 2, 2020.
21. Ian Sutherland , Youngseok Sim, Seul Ki Lee, Jaemun Byun and Kiattipoom Kiatkawsin , Topic Modeling of Online Accommodation Reviews via Latent Dirichlet Allocation, Sustainability 2020, 12, 1821; doi:10.3390/su12051821
22. Ashesh Iqbal ,Sumi Khatun ,Mohammad Shamsul Arefin , andM. Ali Akber Dewan ,ERF: An Empirical Recommender Framework for Ascertaining Appropriate Learning Materials from Stack Overflow Discussions, Computers 2020, 9(3), 57; https://doi.org/10.3390/computers9030057
23. Tian Shi, Kyeongpil Kang, Jaegul Choo, Chandan K. ReddyShort-Text Topic Modeling via Non-negative Matrix Factorization Enriched with Local Word-Context Correlations, DOI: https://doi.org/10.1145/3178876.3186009, WWW '18: Proceedings of The Web Conference 2018, Lyon, France, April 2018.
24. Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. 2016. Topic Modeling for Short Texts with Auxiliary Word Embeddings. In Proceedings of the 39th International ACM

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication：Online Open Access
Vol: 40 Issue: 09-2021
DOI 10.17605/OSF.IO/ F6JVP

SIGIR conference on Research and Development in Information Retrieval . ACM, 165–174.

25. Kamel Alrashedy ∗, Dhanush Dharmaretnam , Daniel M. German , Venkatesh Srinivasan , T. Aaron Gulliver, SCC++: Predicting the programming language of questions and snippets of Stack Overflow, The Journal of Systems and Software 162 (2020) 110505.
26. Guilherme Raiol de Miranda12, Rodrigo Pasti2, and Leandro Nunes de Castro12, Detecting Topics in Documents by Clustering Word Vectors, Distributed Computing and Artificial Intelligence ,DOI: 10.1007/978-3-030-23887-2_27, January 2020.