ISSN 1533-9211

DOI 10.17605/OSF.IO/6DT3K

A NOVEL APPROACH ON TOPIC MODELING MODELS THROUGH DISTRIBUTED FILE SYSTEM IN SOCIETAL COMMUNICATION EMPOWERMENT FOR JOB ASSISTANTSHIP

K. PUSHPA RANI[1], PELLAKURI VIDYULLATHA[2] and Dr. K. SRINIVAS RAO[3]

[1]Research Scholar, CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.
[2]Professor, CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.
[3]Professor, Department of Computer Science and Engineering, MLR Institute of Technology, Hyderabad.

Abstract

To research public opinion on technical terms or topics, the present research programme uses data from social networks. For those looking for employment, the perception of technical phrases or subjects by the general public and their effects on the environment and society are crucial issues. For legislation and the implementation of mitigation programmes, public aid is also crucial. For a better knowledge of the social environment and social dynamics, public opinion study is crucial. Since social media data offers incredibly valuable information on public attitudes and responses to conflicting socio-technical terms or problems from various perspectives, such as quorum, stack overflow, and Yahoo!, it is one of the many sources of public opinion data that is of particular interest to researchers. It responds to Twitter, among other things, and is frequently used to track and assess how society responds to a natural or societal anomaly. In order to identify a variety of topics in the topic templates, data on social media is typically acquired by searching for keywords or a specific topic. In conventional topic models, users can provide an inaccurate number of topics, which leads to subpar grouping outcomes. In these situations, accurate representations are crucial for retrieving data and identifying cluster trends. In order to solve a current issue, viable methods for modelling themes are related to unclassified and incorrect texts or topics. These techniques are the Distributed Latent Semantic Analysis (DLSA) and the Distributed Latent Dirichlet Allocation (dLDA). This document provides a brief overview of the country's public question-and-answer system and traces the evolution of significant problems and initiatives, paying particular attention to the automatic dissemination of pertinent customer feedback and knowledge of pertinent awareness-raising information you seek. Opportunities for housing and employment for the newest technologies in global empowerment. Finally, topic models outperform existing models in terms of precision for obtaining more pertinent responses from a placement and interview perspective, according to the experimental findings.

Keywords: F-Score, Hadoop, LSA, LDA, overflow, Quora, Topic models, stack, Twitter API

1. INTRODUCTION

Often, software engineers and programmers search for answers to questions using websites like Stack Overflow. Analyzing this data [1-3] can give an idea of what aspects of programming and APIs are the most difficult to understand. In this article, to classify overflow stack problems into two overlapping views that is programming concepts and the type of information sought. The overflow attack contains a large amount of information related to a wide range of topics in computer programming. The user posting the question should indicate the category of the question by selecting specific tags that the user deems most appropriate for the context of the question. The user can select the tag from a list of existing tags that other users have already defined, or they can choose to define a completely new tag. It is recommended that the user always prefers the existing tags and creates a new one only when

ISSN 1533-9211
DOI 10.17605/OSF.IO/6DT3K

they consider that there is strong evidence that the problem, they post covers a new topic that no one else has asked before on this site. This allows you to quickly retrieve answers to future questions. However, the poster in question generally indicates multiple bookmarks and there is currently no way to prioritize these bookmarks. Also, the tags available in Stack Overflow may not be appropriate or too general to contextualize the question, which can lead to inappropriate requests to answer future questions. "Because users assign the label, they are also subjective and open to user interpretation [4-6]". Various publications have been made using in-stream stack data exploring a wide variety of topics. Some of these topics that are relevant to our current study include, but are not limited to, finding expert users in response to community questions, trying to find faulty project documentation through topic analysis, answering questions about unanswered stack overflow questions, Stack Overflow Label Prediction and various others that can be found in quotes. Although finding the right answers and improving the quality of your labels is an ongoing effort, the excessive number of responses and click-through rates to achieve a valid response is another challenge. Stack Overflow is a search engine with answers to a non-personalized question and can lead to slow access to information. "The purpose of our study is to present the user with the 10 most relevant questions, such as the consumer inquiry [7]". To accomplish this by developing a k-mean classification model that labels the question with the most appropriate marker or context. Then use topic and ensemble modeling techniques to find similarities between user queries and corpus questions.

2. BACKGROUND KNOWLEDGE

In this section, after the 1960s, "research on artificial intelligence has just begun, scientists have suggested that computers (A) answer questions (Q) using natural language processing (NLP), it can be considered as a Rudimentary response system". Question and answer systems became fashionable for some time in the field of natural language in the 1980s. However, "with the growth of large-scale word processing technology, research on response systems is lacking [8-9]". In previous years, with the rapid development of networks and IT (information technology), people's desire to receive information faster again encouraged the development of response systems. Companies and research institutes participate in this development, including Microsoft, IBM and MIT (Massachusetts Institute of Technology). In 1999, TREC (Text Retrieval Conference) introduced an automatic response to project tracking questions. Since then, the Q&A track has gradually become one of TREC's most popular projects. "Many countries have developed several relatively mature question-and-answer systems. The open question and answer software systems have a Start system developed by the InfoLab group of the Laboratory for Computer Science and Artificial Intelligence at MIT [10-13]". However, the question-and-answer systems used to solve assignments for a particular course are very rare. As a result, an intelligent question-and-answer system has been developed that returns answers to users' questions according to the principles of a course-based course.

ISSN 1533-9211

DOI 10.17605/OSF.IO/6DT3K

3. DATA COLLECTION

The dataset was downloaded from social sentiment analysis and the official view of the Stack Overflow blog. The command includes information from about 2.98,000 individuals. Every year, Stack Overflow runs a survey to learn more about the interests of its users. Based on the responses, different analyses can be performed with the data. At first, you attempted to download the file's Stack Overflow dataset. The data file, nevertheless, had a zip and was 12 gigabytes large. "Due to the limitations of our local computer setup and the inability to work with such a large dataset, we chose to use the dataset provided by Kaggle in this regard". Kaggle contains a specific dataset for python problems. Hadoop is open source and supports the processing storage of extremely large data corpus in an hdfs. Map reduction is also a processing and programming the model for Java, python, R and spark based distributed calculations. The Map Reduce algorithm contains two important tasks, namely Map and Map Reduct. The card accepts one data set and converts it to another data set, where the individual elements are divided into tuples (key / value pairs).

Table 1: Memory Segregation in Hadoop ecosystem

Data sets	File type	(File size)	(Block size)	Modification Time	File Permissions	Owner (O)
	Directory			12/7/202216:45	"rwxr-xr-x"	pushpa
1	File	6.86	128 MB	12/7/202216:47	"rw-r--r—"	pushpa
2	File	6.86	128 MB	12/7/202216:48	"rw-r--r—"	pushpa
3	File	156	128 MB	12/7/202216:50	"rw-r--r—"	pushpa
4	File	156	128 MB	12/7/202216:54	"rw-r--r—"	pushpa
5	File	156	128 MB	12/7/202216:57	"rw-r--r—"	pushpa

4. DATA PRE-PROCESSING

One of the key ideas before applying NLP and text analysis approaches is text preprocessing since the elements extracted from the text be they symbolic sentences, words, or phrases serve as the foundation for subsequent analysis. To increase the accuracy of the classifiers, the text must be deleted, standardized, and more text information must be gleaned from the notes.

First, clean up the content by eliminating any html tags, stop words, and extraneous keywords. The regexp tokenizer class from the NLTK framework, which tokenizes words based on regular expression-based models to partition sentences and subsequently remove those words from sentences, should be used for word tokenization. Additionally, case-sensitive conversion needs to be done. Labels must be normalized in order to produce a clear, uniform version of the labels.

Spelling and stem corrections as well as stemming were employed as cleaning methods for the markers. "There are several models to choose from when using feature recovery techniques, including Word2vec, Bag of Words, and TF-IDF [15–17]". Test the Word Bag model, which converts the text into a vector that indicates the frequency of all the individual words contained in the text vector space for that text, after starting with the vector space model, in which text data is represented as numerical vectors of terms for vector dimensions.

ISSN 1533-9211

DOI 10.17605/OSF.IO/6DT3K

It occurs more frequently and causes eclipses and therefore, "they might not happen as frequently. The term document frequency and inverse frequency (TF-IDF) [18]" was then put to the test. "Create the vector using the TF-IDF [19–20] weight and the body of the question serves as the primary attribute from which we derive the function [21–22]". This framework addresses Hadoop distributed topic modelling strategies to improve procedures and outcomes.

5. METHODOLOGY

Our survey replied to the user query by first giving the question the proper label or intent and, in the instance of Stack Overflow, matching the user query with the question that was most like it. The first ten most crucial questions were then provided. The study's first section employs two distinct methodologies: Using genetic markers in Python, k-means grouping documents and modelling topics to group inquiries. Topic modelling groups topics using probabilistic models, whereas clustering employs surprising clustering methods.

A body of unstructured documents is used as the input for topic modelling, a text extraction technique that produces the best-classified terms in a topic and documents that are linked to it. Comparatively to document classification, where only one subject is connected to text, a single text document can be associated with several topics using a topic modelling approach. Instead of creating a collection of texts or documents, topic modelling develops a collection of words, where each word is a combination of various themes, each of which is given a particular weight.

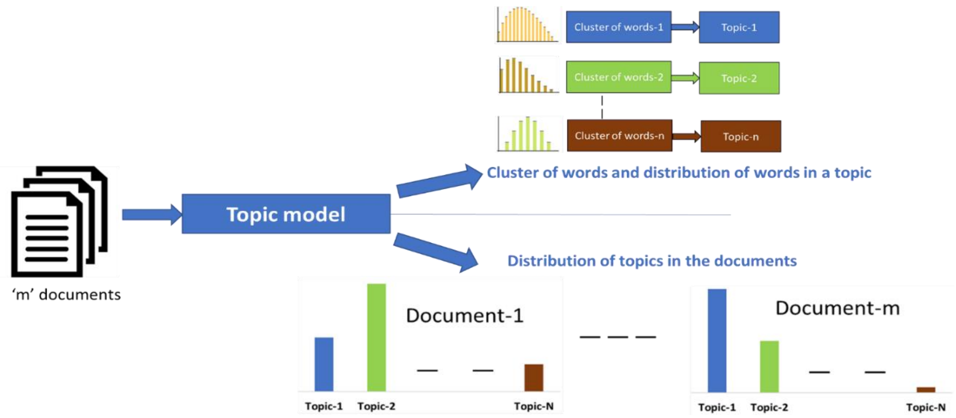

Fig 1: Topic Models

By using topic modelling, latent themes in a set of papers are automatically uncovered. Finding the group of terms from the provided content is done using an unsupervised text analytics technique. This collection of words serves as the topic. It is possible for a single document to be connected to several themes.

ISSN 1533-9211

DOI 10.17605/OSF.IO/6DT3K

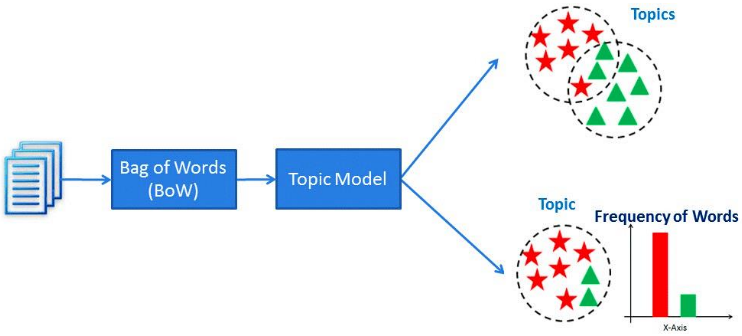

Fig 2: BOWS

A text document or article is classified into a predefined set of classes as part of the supervised machine learning task known as text classification. Finding clusters of words that frequently occur together in text texts is a process known as topic modelling. "Topics" are made up of these clusters of words with a common theme. Since it is an unsupervised learning method, the collection of potential topics is unknowable. To address the issue of text classification, topic modelling can be utilized. While text classification divides the text into several classes, topic modelling identifies the subjects that are presented in a document.

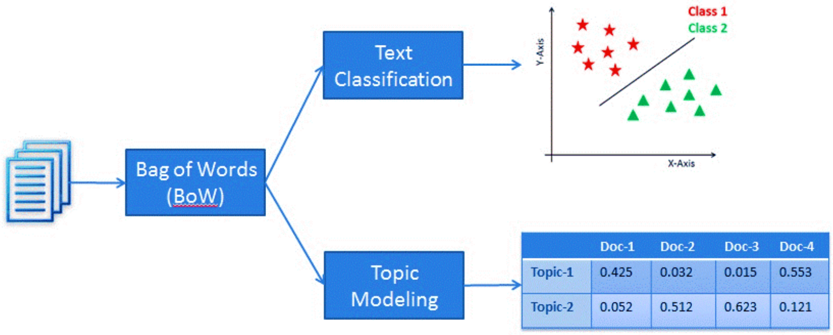

Fig 3: Comparison between Text Classification and Topic Modeling

The Distributed Latent Semantic Analysis:

LSA is frequently referred to as distributed latent semantic analysis (DLSA) or distributed latent semantic index (DLSI), and it employs the bag of words (BoW) model, which creates a term-document matrix (occurrence of terms in a document). Documents are represented by columns, and terms by rows. DLSA can discover latent themes by performing a matrix decomposition on the document-term matrix using singular value decomposition. The technique of dimension reduction or noise reduction is widely used in DLSA.

ISSN 1533-9211
DOI 10.17605/OSF.IO/6DT3K

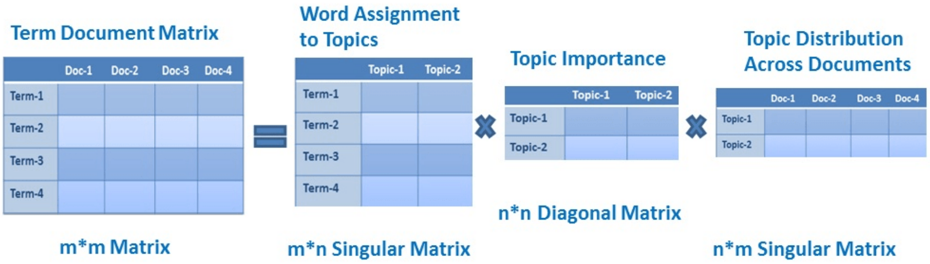

Fig 4: Term Document Matrix

The topic coherence metre is a popular tool for assessing topic models. The latent variable models are employed. Each topic that is generated has a list of terms. The average/median pairwise word similarity scores of the terms in a subject are found in the topic coherence measure. A good topic model will be one with a high topic coherence score model value.

One of the key methods for modelling themes is latent distributed semantic analysis, or dLSA. The main concept is to create a topic-topic matrix out of a matrix of terms and documents. One of the key methods used in topic modelling in the environment is called Distributed latent semantic analysis (dLSA) and the creation of our document matrix comes first. We may create a m n A matrix, where each row corresponds to a document and each column corresponds to a word, using the m documents and n words in our dictionary.

The simplest LSA can merely count the number of times the word j appears in document I for each entry. The standard number does not, however, perform well in practise since it does not consider the meaning of each word in the document. For instance, the term "nuclear" tells us more about the topic(s) of the document than the word "evidence." As a result, the dLSA model frequently replaces the raw integers in the document matrix with a tf-idf result. The Tf-idf, sometimes referred to as the inverse frequency of the document, establishes the weight of the phrase j in document I.

$$W_{i,j} = Mul\ \{tf_{i,j}, \log(N/df_j)\}$$

Distributed Latent Dirichlet Allocation:

Distributed latent Dirichlet allocation (dLDA), a generative statistical model used in natural language processing, enables the explanation of sets of observations by unobserved groups that explain why some portions of the data are similar. For instance, if observations are words that are gathered into documents, it is hypothesized that each document is a combination of a few different subjects, and that each word's origin can be linked to one of those topics.

The LDA model is visually demonstrated in the figure below. The model's objective is to identify the subject and document vectors that explain the various documents' initial "bag-of-words" representation.

ISSN 1533-9211
DOI 10.17605/OSF.IO/6DT3K

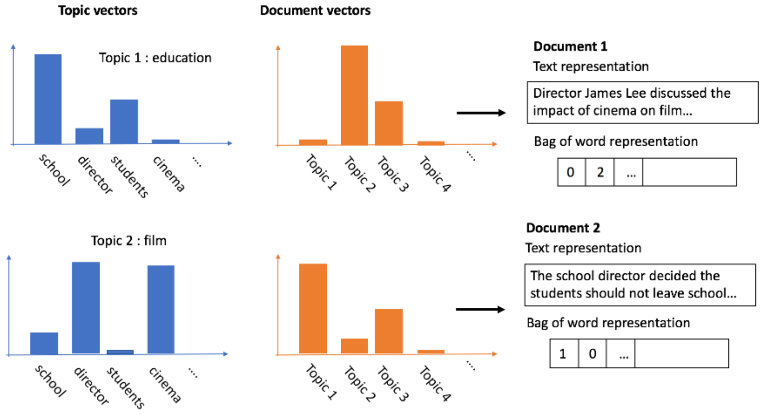

Fig 5: dLDA

It is crucial to note that you are relying on the presumption that the topic vectors will be understandable because, in the absence of this, the model's output is essentially useless. You are essentially trusting that, given enough data, the model will identify the terms that frequently occur together and group them into discrete "themes."

A straightforward probabilistic model with good performance is LDA. The document vectors frequently show the pattern and structure in documents because they are sparse, low-dimensional, and easy to read. The number of subjects that appear in the collection of papers must be accurately estimated. Additionally, each subject vector must have a unique nominator "topic" assigned manually. LDA may experience the same drawbacks as the bag-of-words model since both are employed to represent the documents. Without considering any structure or the way that these words interact locally, the LDA model learns a document vector that predicts the words that will appear in that document [23-24].

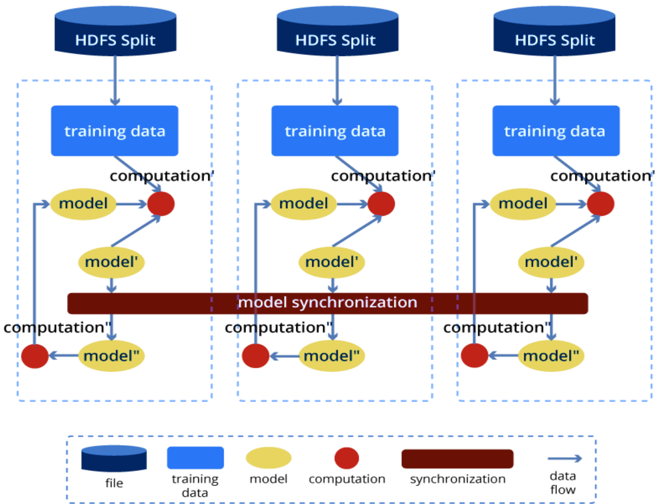

Fig 6: HDML-LDA

The Seybold REPORT

ISSN 1533-9211

DOI 10.17605/OSF.IO/6DT3K

In the initialization phase of training, Harp will load the local data split on each node into memory, eliminating the need for future disc I/O to access the training data. By default, Hadoop Map Reduce uses the data splitting strategy it supports.

Hadoop offers distributed dataset abstractions, group communication, and synchronization methods for model data. Since the fundamental computation of machine learning algorithms is the model update, parallelizing the core computation of the model update causes issues with model consistency and synchronization. For networked machine learning applications, Harp's distinctive abstractions based on collective synchronization mechanism are advantageous in terms of expressiveness, efficiency, and effectiveness.

Evaluation Performance:

In this work, topic modelling techniques were employed to locate similar questions, and several similarity vectorization models were compared before the most pertinent questions were effectively obtained. To define the 20 most pertinent questions for each user query, ensemble models, topic modelling, and similarity for k themes were used to quantify the importance of the similarity of the questions. The total number of issues in the corpus is used to define the suitable document.

A question that is related to the specific document is the definition of the received document. The distributed modelling approaches strike a balance between solving the current issues and obtaining accurate results in table 7. Distributed LDA of the topic equilibrium model generated the best results in the comparative analysis shown in figs. 10 and 11. It had superior accuracy in the estimation of related indicators.

Table 2: Document Topic Count Matrix

Document /topics	Topic.1	Topic.2	Topic.3	Topic.4	Topic.5	Topic.6	Topic.7	Topic.8	Topic.9	Topic.10
Doc-1	0	1	1	1	1	2	0	1	0	1
Doc-2	0	1	1	1	1	1	0	0	0	2
Doc-3	0	2	0	0	1	1	0	2	0	0
Doc-4	1	0	2	1	0	0	1	2	1	0
Doc-5	1	2	1	1	1	0	2	0	0	0
Doc-6	0	0	1	4	1	1	0	0	1	1
Doc-7	0	0	0	1	1	0	3	1	2	1
Doc-8	1	0	0	0	1	2	1	2	0	2
Doc-9	0	1	0	1	0	1	0	1	0	1
Doc-10	5	0	1	3	3	2	5	1	2	0

In table 2, which generates a matrix of phrases for documents (occurrence of terms in a document). Terms are represented by rows, and documents by columns. By employing Singular value decomposition to perform a matrix decomposition on the document-term matrix, DLSA may learn latent themes. DLSA is frequently applied as a technique for dimension reduction or noise reduction.

ISSN 1533-9211

DOI 10.17605/OSF.IO/6DT3K

Table 3: Topic Probabilities per Document

Document /topic	Topic.1	Topic.2	Topic.3	Topic.4	Topic.5	Topic.6	Topic.7	Topic.8	Topic.9	Topic.10
Doc-1	0.06	0.24	0.12	0.06	0.06	0.12	0.12	0.12	0.06	0.06
Doc-2	0.06	0.18	0.12	0.06	0.06	0.18	0.12	0.12	0.06	0.06
Doc-3	0.06	0.19	0.13	0.06	0.06	0.13	0.13	0.13	0.06	0.06
Doc-4	0.06	0.06	0.29	0.06	0.06	0.12	0.12	0.12	0.06	0.06
Doc-5	0.10	0.06	0.14	0.05	0.05	0.14	0.10	0.14	0.19	0.05
Doc-6	0.07	0.07	0.13	0.07	0.07	0.07	0.07	0.13	0.20	0.13
Doc-7	0.06	0.06	0.13	0.13	0.06	0.06	0.06	0.19	0.19	0.06
Doc-8	0.19	0.05	0.10	0.19	0.05	0.19	0.05	0.05	0.05	0.10
Doc-9	0.12	0.06	0.12	0.18	0.06	0.06	0.06	0.12	0.18	0.06
Doc-10	0.11	0.17	0.09	0.04	0.02	0.09	0.04	0.04	0.04	0.09

Table 3 shows the results of a probabilistic topic model analysis of document content and topic meanings.

Table 4: Word Probabilities per Document

Document word	Word-1	Word-2	Word-3	Word-4	Word-5	Word-6	Word-7	Word-8	Word-9	Word-10
Doc-1	0.059	0.235	0.118	0.059	0.059	0.118	0.118	0.118	0.059	0.059
Doc-2	0.059	0.176	0.118	0.059	0.059	0.176	0.118	0.118	0.059	0.059
Doc-3	0.063	0.188	0.125	0.063	0.063	0.125	0.125	0.125	0.063	0.063
Doc-4	0.059	0.059	0.294	0.059	0.059	0.118	0.118	0.118	0.059	0.059
Doc-5	0.095	0.048	0.143	0.048	0.048	0.143	0.095	0.143	0.190	0.048
Doc-6	0.067	0.067	0.133	0.067	0.067	0.067	0.067	0.133	0.200	0.133
Doc-7	0.063	0.063	0.125	0.125	0.063	0.063	0.063	0.188	0.188	0.063
Doc-8	0.190	0.048	0.095	0.190	0.048	0.190	0.048	0.048	0.048	0.095
Doc-9	0.118	0.059	0.118	0.176	0.059	0.059	0.059	0.118	0.176	0.059
Doc-10	0.174	0.174	0.087	0.043	0.217	0.087	0.043	0.043	0.043	0.087

Table 4 shows the results of an analysis of word meanings and document content using several probabilistic topic models.

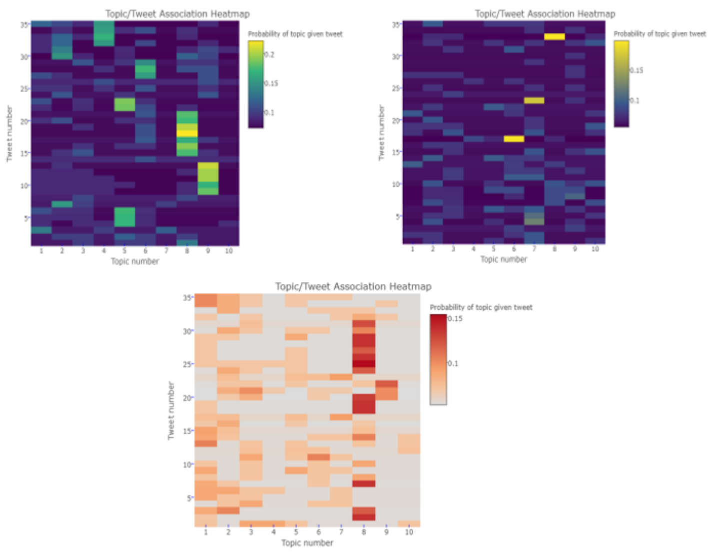

Fig 7: Probability of topic Distribution

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/6DT3K

The graphical examination of topic distribution probability in the Hadoop ecosystem and several probabilistic topic models have been used to assess document content and word meanings.

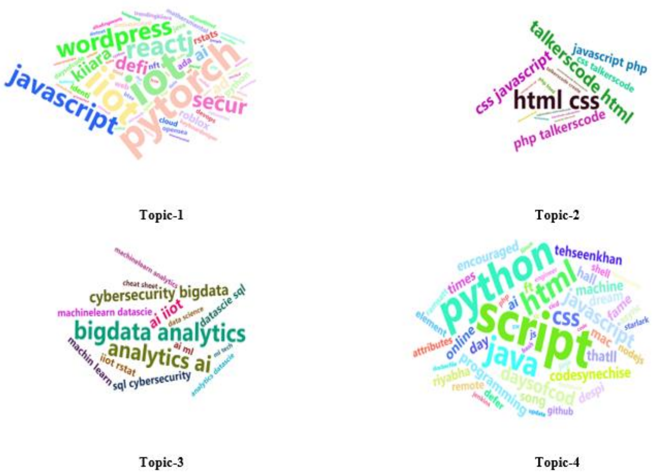

Fig 8: Induvial Topic Cluster

The graphical examination of topic cluster distribution probability in the Hadoop environment and several probabilistic topic models have been used to assess document content and word meanings.

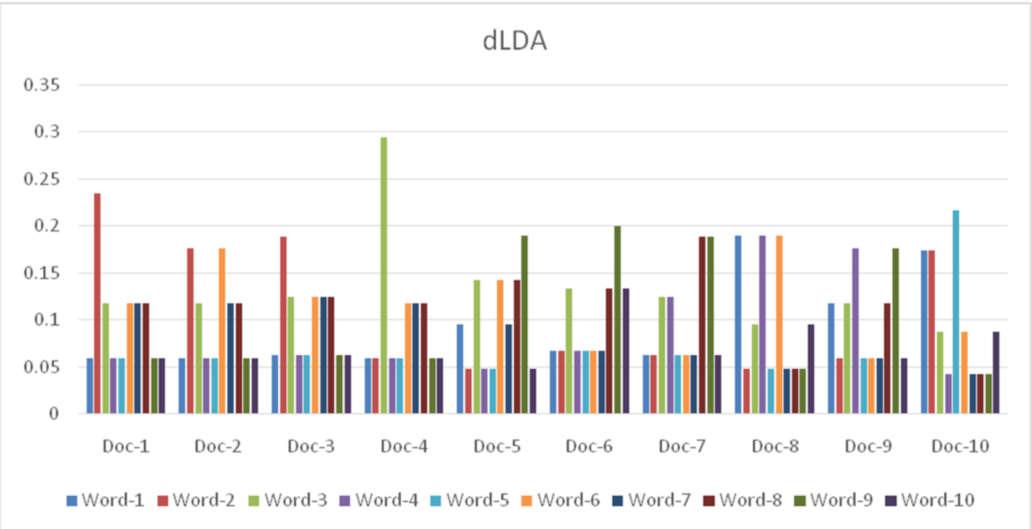

Fig 9: Topic Cluster

Several probabilistic topic models have been used to analyze the content of documents and word meanings, and the results are shown in fig. 9. The distributed LDA is supervised machine learning approach addressed best results.

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/6DT3K

Table: 5 Comparison Results

Methods	pre	acc	rec	f1	Topics
dLDA	0.87	0.89	0.77	0.82	50
dLSA	0.12	0.67	0.54	0.61	50
dLDA	0.93	0.94	0.87	0.90	100
dLSA	0.67	0.71	0.74	0.69	100

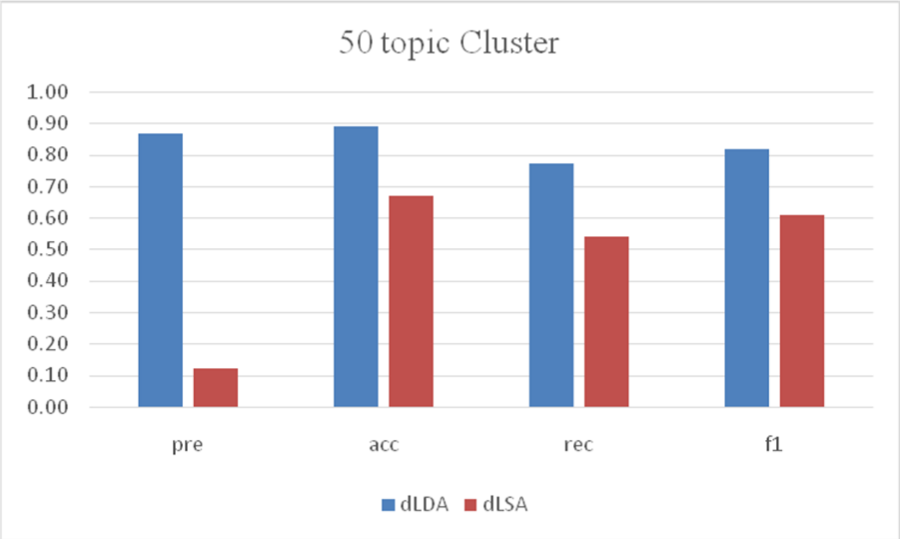

Fig 10: Comparison Results with 50 topic cluster

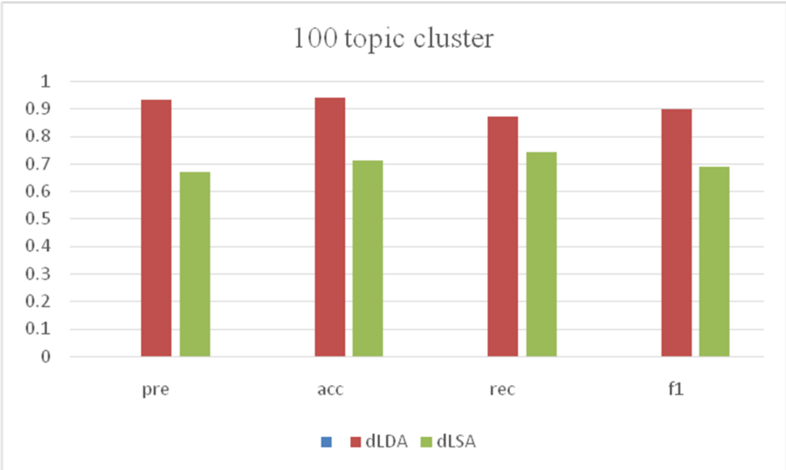

Fig 11: Comparision Results with 100 topic cluster

Table 5 and Figures 10 and 11 in this section addressed model results. Distributed LDA is a supervised machine learning technique that addressed best outcomes and provided a graphical analysis of topic distribution probability in the Hadoop ecosystem when compared to dLSA.

ISSN 1533-9211

DOI 10.17605/OSF.IO/6DT3K

Several probabilistic topic models have been used to assess document content and word meanings.

7. CONCLUSION

This study looks at how topic models provide insight on programming (P) languages and issues that experts encounter. We were able to draw conclusions from the question type analysis that would not have been achievable otherwise. They were able to show that the types of inquiries are the same across programming languages and to give a method for determining which types of queries are primarily related to structural structure identifiers. Topic modelling creates a representation of the collection of text documents in the topic space because it gives themes that are present in every text document.to achieve an overall performance by combining the topic modelling and additionally, topic modelling utilizing the dLDA algorithm and the similarity measure produces superior results. Future improvements will include more attributes (topics) to analyse questions and answers and categorise the user authorization system, better time and space distribution of the models to get a better location for user displacement, and platform-related advanced technologies from the users who are talking in the Twitter, Stack Overflow, and Quora databases.

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, 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.

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/6DT3K

11) Peng Zhang, Department of Mathematics, Zhejiang University, Hangzhou, 310027 China ; WanhuaSu Statistical inference on recall, precision and average precision under random selection,2012

12) Edi Surya Negara, Dendi Triadi, Ria Andryani,Topic Modelling Twitter Data with Latent

13) Dirichlet Allocation Method,DOI: 1109/ICECOS47637.2019.8984523,,Electronic ISBN: 978-1-7281-4714-7,Print on Demand(PoD) ISBN: 978-1-7281-4715-4,2019

14) Pablo Ormeño ,Marcelo Mendoza , and Carlos Valle ,Topic Models Ensembles for AD-HOC Information Retrieval Information 2021, 12(9), 360; https://doi.org/10.3390/info12090360

15) Sung-Hwan Kim and Hwan-GueCho ,User–Topic Modeling for Online Community Analysis,Appl. Sci. 2020, 10(10), 3388; https://doi.org/10.3390/app10103388

16) Dhelim, S.; Aung, N.; Ning, H. Mining user interest based on personality-aware hybrid filtering in social networks. Knowl. Based Syst. 2020, 206, 106227

17) Baechle, C.; Huang, C.; Agarwal, A.; Behara, R.; Goo, J. Latent topic ensemble learning for hospital readmission cost optimization. Eur. J. Oper. Res. 2020, 281, 517–531.

18) Pourvali, M.; Orlando, S.; Omidvarborna, H. Topic Models and Fusion Methods: A Union to Improve Text Clustering and Cluster Labeling. Int. J. Interact. Multimed. Artif. Intell. 2019, 5, 28–34

19) Fiandrino, S.; Tonelli, A. A Text-Mining Analysis on the Review of the Non-Financial Reporting Directive: Bringing Value Creation for Stakeholders into Accounting. Sustainability 2021, 13, 763.

20) yerragudipadusubbarayudu, AlladiSureshbabu, “Distributed Multimodal Aspective on Topic Model using sentiment analysis for Recognition of Public Health Surveillance” Expert Clouds and Applications, ISBN: 978-981-16-2126-0, https://link.springer.com/chapter/10.1007/978-981-16-2126-0_38

21) Ammirato, S.; Felicetti, A.M.; Raso, C.; Pansera, B.A.; Violi, A. Agritourism and Sustainability: What We Can Learn from a Systematic Literature Review. Sustainability 2020, 12, 9575.

22) Farkhod, A.; Abdusalomov, A.; Makhmudov, F.; Cho, Y.I. LDA-Based Topic Modeling Sentiment Analysis Using Topic/Document/Sentence (TDS) Model. Appl. Sci. 2021, 11, 11091.

23) P. Vidyullatha, P. venkateswara Rao, Big data sentimental analytics on social media using Rhadoop-Hive, "Materials Today: Proceedings", January 2021.

24) Devisetty, S.D.P., Sai, Y.M., Yadav, A.V., Vidyullatha, P., “Sentiment analysis of tweets using rapid miner tool “, International Journal of Innovative Technology and Exploring Engineering, 2019, 8(6), pp. 1410–1414