Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 7s (2025)

Effective Information and Extraction an Improved Semantic Pattern Based Approach in LDA Model using Bigdata

Dr. K. Sundravadivelu[1]

[1] Assistant Professor, Department of Computer Science, Madurai Kamaraj University, Madurai, Tamil Nadu, India.
Email: svadivelu2021@gmail.com[1]

Article History:

Received: *25-10-2024*

Revised: *03-12-2024*

Accepted: *12-12-2024*

Abstract:

Big text data content is produced by more than ninety percent of online sources. The data occurs primarily in two forms, structured and unstructured, depending on the representation schemes underlying the devices or platforms that generate them. The focus of this research is on the unstructured text data, as it provides for unforeseen opportunities in decision-making tasks and wider scope for analytics. The proposed work attempts to mitigate the complexity by combining semantic analysis with content modelling. To achieve the data model, Latent Drichlet Allocation topic modelling technique is improved using Frequent Sequential Pattern mining algorithm. Hierarchical topic clusters in documents are viable to be extracted with the help of LDA when word associations are modelled using frequent and sequential patterns. The semantic patterns obtained from such processes are used for knowledge discovery. This knowledge can be used for other similar mining tasks such as classification, clustering or information extraction. This study has four primary contributions such as, to uncover patterns in unstructured text content, semantic information extraction, pattern improvements, topic modelling, and cluster assignments. The proposed model is compared with that of the three other existing alternatives in this regard, SLDA (Sentiment LDA), HME-LDA (Hierarchical clustering Max-Ent) and FP-LDA (Frequent Pattern). The experimental results prove that the proposed model outperformed existing models with a considerable margin.

Keywords: Pattern Mining, Sentiment LDA, Frequent Pattern LDA, ISP-LDA, Hierarchical clustering Max-End LDA, etc.

1. Introduction

Widely used in knowledge-driven organizations, text mining is the process of examining large collections of documents to discover new information or help answer specific research questions. Text mining identifies facts, relationships and assertions that would otherwise remain buried in the mass of textual big data [1]. Once extracted, this information is converted into a structured form that can be further analyzed, or presented directly using clustered HTML tables, mind maps, charts, etc. Text mining employs a variety of methodologies to process the text, one of the most important of these being Natural Language Processing (NLP).

Machine learning is an artificial intelligence (AI) technology which provides systems with the ability to automatically learn from experience without the need for explicit programming, and can help solve complex problems with accuracy that can rival or even sometimes surpass humans. However, machine learning requires well-curate input to train from, and this is typically not available from sources such as electronic health records or scientific literature where most of the data is unstructured

text [3]. Concordance is used to recognize the particular context or instance in which a word or set of words appears. Analyzing the concordance of a word can help understand its exact meaning based on context.

This paper proposes two schemes for aspect-based opinion mining. The first scheme is based on the inverted list and the SLDA (SentiWordNet WordNet-Latent Dirichlet Allocation) model proposed in this paper. The second scheme is based on the inverted list and the HME-LDA (Hierarchical Clustering MaxEnt-Latent Dirichlet Allocation) model proposed in this paper. The SLDA model is an optimized LDA model based on the WordNet and SentiWordNet, where the WordNet is for the similarity calculation of words and seed words and the SentiWordNet is for the separation of the opinion target words and the opinion words, while the HME-LDA model is an optimized LDA model based on SLDA and MaxEntLDA [22].

This methodology uses Latent Dirichlet Allocation (LDA), a probabilistic topic-modeling technique to discover the hidden semantic structures from a given textual corpus. The output of this study is a systematic competency map comprising the essential knowledge domains, skills, and tools for big data software engineering. In this study, a semi-automatic methodology was proposed to analyze the content of online BDSE job ads. Because of those reasons, people are interested in preprocessing raw text in a way that reduces its dimension. One way to achieve that is to use LDA [5] in an unsupervised manner. The Boost Multi-class ISP-LDA model, for text classification that is an ensemble of Multi-class LDA models is proposed to solve variation inference.

The research work consists of three main layer such as data layer, pattern layer and topic model layer. Each layer is made up of sub process stages such as Pre-processing text data, Set of Documents, Pattern Discovery, Semantic Information and Extraction, Pattern Improvements, Topic modeling and Cluster assignments.

- In pre-process stage; the data is cleaned for eliminating noisy data. The text is converted into lower case, stop word removal, and punctuations are removed from the text. In the next step, all tokens of length less than three characters and greater than twenty-five characters are removed from a text corpus.
- In the next stage, using sequential pattern mining algorithm, the document is converted into paragraph and further into individual terms for identifying the word sense.
- Pattern Discovery stage involves frequent & closed pattern Techniques, to understand the underlying patterns in the consecutive words as, sometimes system falsely identifies the negative documents.
- Semantic Information & extraction stage uses d pattern algorithm. Patterns are organized based on the weights of terms. It then evaluates the term weights and discovers specific patterns in the set of documents.
- set of documents – no of documents related to various long news stories, blogs, lengthy reviews – no of documents, number of pages, paragraphs, words, etc., Pre-processing natural language processing- pattern identification, pattern inherent, formal informal patterns, pattern extraction , interpretation & evaluation , how to interpret patterns, evaluation of inherent patterns, term weighting and its effect on patterns formed,

- Pattern Improvements are applied to reduce false positive assignments due to the presence of synonym words. It eliminates both noise documents and noise patterns to get positive assignments.
- LDA Algorithm is a topic modelling algorithm used to understand the inherent topics in the document. The LDA model to find out topics for the articles belonging to multiple categories. It is achieved using the similar other textual content, the vocabularies derived from them and then the relationship between its terms.

- The consecutive terms are grouped based on similarity and underlying relationships into multiple clusters using the Hclust algorithm. The overall system architecture is shown in Figure.1.

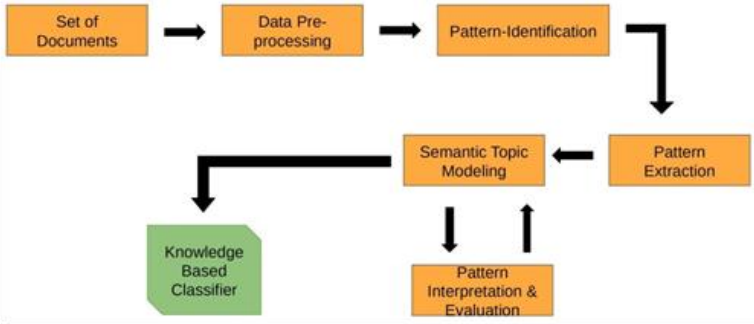


Fig.1. Proposed architecture of Improved Semantic Pattern Based LDA Model (iSP-LDA)

1.1 Data Collection and pre-processing

The data is a collection of research articles published in online databases, spanning acros multiple topics, such as, science, language, arts, etc. The total number of documents taken for the study is 96,720 articles from Google Scholar, Pubmed, NCBI, Elsevier and IEEE, Springer nature books, Science Direct. The metadata about the data collection consists of Author, Date of publication, Subject, Accession number and Publisher details. The Data is then processed for removal of unwanted or noisy text content. Before data cleaning, the documents are reduced to paragraphs and then to words without losing word order [24].

2. Pattern Discovery

2.1 Sequential pattern mining

This is the crucial step in the proposed work. It uses text mining techniques to apply sequential pattern mining algorithms to extract text patterns in the corpus. It helps to discover the word sense, sentence relationships, semantic meanings and ambiguity reduction [14].

This reduces the search space for extraction of patterns representing various topics. The challenge of sequential pattern mining entails extracting frequent non-contiguous or contiguous sub-sequences. Because algorithms must generate and test a combinatorial enormous number of intermediate sub-sequences, this is considered difficult in terms of temporal complexity [22]. Furthermore, most of the retrieved sequences in sequential pattern mining are quite similar to each other and excessively redundant. This task requires less computing power and eliminates the issue of redundancy in the extracted patterns [15].

Pattern discovery extracts the patterns from the corpus where each pattern includes a set of terms. It uses the pattern taxonomy model to extract and update the discovered successful patterns to reveal the information present in the text documents [11]. The taxonomy provides structure to the patterns while conserving semantic meaning and word order of the data. It provides sub-set relations through sequential patterns or words in the corpus. Which T= {t1, t2, t3, t4 ...tz} be a bunch of terms that are extricated from the algorithm. Let T = {t1; t2... tm} be a bunch of terms (or watchwords) that can be removed from the arrangement of positive documents, D+. The negative documents are represented as D- [12].

To extract the frequent patterns the minimum support is deputed. The helped value is calculated based on absolute support ((sup_{a1}) and relative support (sup_{r1}). The outright assistance is the quantity of event of terms (X) in PS1 (D), for example, sup_{a1} (X) = |X| and the general backings are the small portion of sections that contain the pattern.

sup_{r1} (X) = |X| / |PS1 (D)|

An ordered list of terms is called as sequential patterns. Saa = < aa1,aa2, aa3….aa_i>is a subsequence of another sequence Sbb =<bb1, bb2, bb3….bb_i> is called Saa is a subset of Sbb if and only if it belongs to j_1, j_2 ... j_b such that $1 \leq j_1 \leq j_2 \leq j_b \leq j$ and $a_1 = b_{j1} \ldots a_i; = b_{jb..}$ A sequential patterns has super patterns and sub patterns. Here Saa is a subset of Sbb and Sbb refers as super pattern of S.[11].A relative support must be greater than 2 (minimum support), then this sequential pattern is said to be frequent pattern[21].

A frequent sequential pattern X is known as closed if not any super pattern x1 of X such that $sup_{a1}(Xl) = sup_{a1}(X)$.

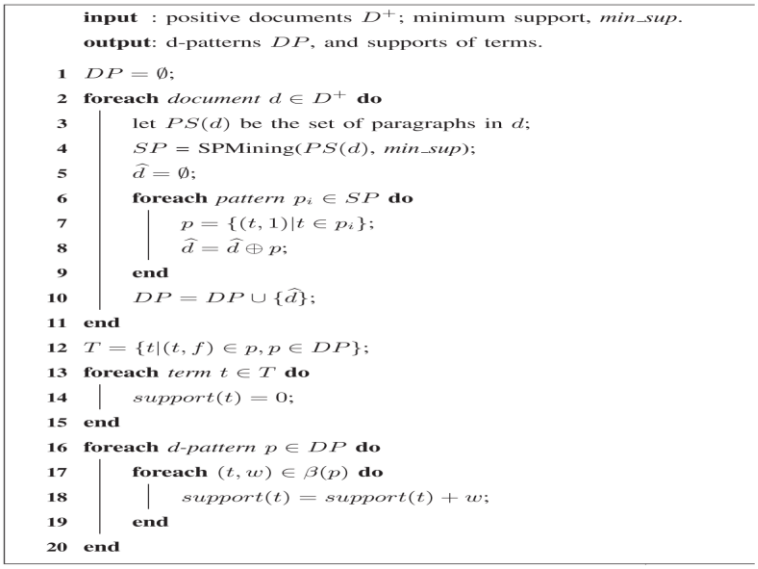
3. Semantic Information and Extraction

The semantic information and extraction, text chunks become data bits, data bits become semantic metadata and semantic metadata become knowledge bytes – data pieces, ready to leverage for insights, decisions and actions Semantic annotation is the process of tagging documents with relevant concepts. The documents are enriched with metadata: references that link the content to concepts, described in a knowledge graph. This makes unstructured content easier to find, interpret and reuse [12].

3.1 D-Pattern Mining Algorithm

A set of d-patterns is used to find out different patterns in every positive document in D-pattern mining algorithm [20]. The deploying process plays a significant role in finding the d-patterns and support assessment terms. Example for Sub-sequence is a sequence Saa= Saa1, Saa2, Saa3….. Saan is a sub sequence of another sequence terms such as Sbb= Sbb1, Sbb2, Sbb3….Sbbm, which was denoted by α, β. A set of integers such as $1 \leq ii1 < ii2 < \ldots < iin \leq m$ such that Sa1=Sb1, Sa2=Sb2……San=Sbm.

Algorithm 1: D-Pattern Mining.



input : positive documents D^+; minimum support, min_sup.
output: d-patterns DP, and supports of terms.

$DP = \emptyset$;
foreach *document* $d \in D^+$ **do**
  let $PS(d)$ be the set of paragraphs in d;
  SP = SPMining($PS(d)$, min_sup);
  $\widehat{d} = \emptyset$;
  foreach *pattern* $p_i \in SP$ **do**
    $p = \{(t, 1) | t \in p_i\}$;
    $\widehat{d} = \widehat{d} \oplus p$;
  end
  $DP = DP \cup \{\widehat{d}\}$;
end
$T = \{t | (t, f) \in p, p \in DP\}$;
foreach *term* $t \in T$ **do**
  $support(t) = 0$;
end
foreach *d-pattern* $p \in DP$ **do**
  foreach $(t, w) \in \beta(p)$ **do**
    $support(t) = support(t) + w$;
  end
end

All found patterns in a positive document are made into d-pattern provide the service a lot of d-patterns stages 6 to stages 9. From that point forward, from stages 12 to 19, term upholds are determined identified with the ordinary structures for all terms in d-patterns. In Algorithm 1, (D pattern Mining) reveals describes the training process which are used to find the set of d-patterns. The main goal of the deploying process is to find out the d-patterns which are available in text documents and also for finding word support assessment. In these words supports are determined on the basis of all the words present in the d-patterns. A set of words or keywords represented as td= {td1, td2…tdn} and a sequence is represented as ss=ss1, ss2…ssn where ssi belongs to the text documents as td.. The d-pattern effectively discovered and term support evaluations are main task for finding deploying process. In Algorithm 1, a positive document consists of d-patterns discovered patterns in a positive report. These are represented as a bunch of d-designs .These d-designs are registered for all terms in d-designs which are dependent on ordinary structures. [13]. the deployed pattern is discovered patterns effectively and they are grouped up. The d-pattern algorithms are used for finding all patterns in certain documents is gathered. Patterns are organized in specific format it evaluates term weights & discovered specific patterns. The d-pattern terms are processed by upholds [19]. The evaluation of Term support is determined by the weight of the term. The semantically rich patterns are extracted in this manner.

3.2 Topic Modeling

The obtained patterns are incorporated in a topic modeling algorithm to extract topics in unstructured manner.

The topic modeling part uses LDA (Latent Dirichlet Allocation) model which is extended by sampling weights from a Dirichlet distribution, the conjugate prior to the multinomial distribution. Depending on the frequency of terms in each cluster the topics are classified using LDA [18]. The aim is to extract topics from research papers and use it for classification of research articles. The topic terms are obtained using cluster analysis, D-pattern and Inner pattern evolving algorithms.

The usage of pattern mining algorithms with LDA will help to extract semantic information from texts as it represents term contextual information. This was lacking in LDA in configure ration of the subsequent item terms and points. This extension allows the model to assign probabilities to data outside the training corpus and uses fewer parameters, thus reducing over- fitting.

In LDA first start with assuming ‘T’ topics in the document. The circle through ‘D’ the document and distribute each word in the report of any of the 'T' topics**.** In step 7 ad 8 for each document loop through each word w and compute $p(w_j|t_k)$ and $p(t_k|d_i)$.after completing the calculation then follow the step 9 for update the $p(w_j|t_{ki}d)$ such as $p(w_j|t_{ki}d_i) = p(t_k|d_i) \times p(w_j|t_k)$, until the loop through each word in each document. In step 10 reassign the topic for currently selected word based on p $(w_j|t_{ki}d_i)$.other repeat for all process in the entire document, finally collected the model. The model is evaluated on known text samples, and validated. The model sometimes, falsely assign a positive pattern for a topic as negative. To solve the issue, pattern improvement module is used.

3.3 Pattern Improvements:

Inner Pattern Evolving: The IPE procedure is utilized to decrease the results of noise patterns. PTM Algorithm is used for finding d-patterns in the positive documents(D+) which are based on minimum support and d-patterns are deployed for find out the term support[13]In the test phase, it calculates the weights for all incoming documents it can be sorted by based on weights. To discovered the patterns are much more specific than all documents. Now and then, the framework falsely lessened negative document as a positive document and diminish the noise documents.

Construct conditional - Frequent Pattern tree from each pattern base. Conditional Frequent Pattern - trees and grows frequent patterns obtained so far. If the conditional Frequent Pattern -tree contains a single path, simply enumerate all the patterns [15]. The Improved IPE will be applicable to the low frequency patterns of text mining for the clients **.**The consideration of further developed internal pattern evolving has been demonstrated to expand accuracy that is the undesirable impacts of regular item-set mining methods have been disposed of significantly by the proposed framework.

3.4 Hclust algorithm: The agglomerative clustering works in a ‘bottom-up’ manner. Each object is initially considered as a single-element cluster (leaf). At each step of the algorithm, the two clusters that are the most similar are combined into a new bigger cluster nodes[16]. The Agglomerative Hierarchical Clustering is the most common type of hierarchical clustering used to group objects in clusters based on their similarity. It’s also known as AGNES (Agglomerative Nesting).

In above algorithm, step1 for preparing the input data in all documents, in step 2 loop the pair of cluster p1 and p2 then merges the clusters. In step4 computing similarity information between every pair of clusters in the data set. Repeat the step2 and step3, until a single cluster containing the group of all documents. In step 6 determining the most commonly used a threshold value. The patterns that make up each topic will be clustered according to their sematic similarity scores. It forms the data dictionary for future classification tasks. The dictionary is again fed into the Topic model for validation. The steps are repeated until optimal accuracy is obtained.

4. Experimental Results

4.1 Datasets:

Reddit submission corpus: - subsetted from master dataset from reddit user submitted data, amounting to 40,000 entries related to health ailments. The submissions that had longer text content were used for the analysis [citation_needed]. Yelp dataset for academia with 25,000 entries related to topic based content is included as another comparison dataset. Amazon product reviews that share elaborate review for various products summing up to 15,000 entries were included for comparison. BBC news article dataset with 30,000 entries was included, to introduce diversity in the topics for the proposed model. All the datasets were chosen keeping the real world content and language as criteria. The Intel Core (TM) i7 with Programming with Tool-R and Windows 10 operating systems with an 8 GB RAM capacity have been employed to implement the system performance a total of TB of dataset. The model was tested in these four datasets and performance was measured with respect to Precision, Recall, F1-score and LDA specific metric, coherence [19].

BBC news articles dataset:

One table with ‘accuracy’ value for these four datasets with respect to four models for before improving patterns and after pattern improvements.

4.2 **Accuracy:**

Accuracy is calculated as: Accuracy is one of the performance measures and it is simply a ratio of correctly predicted observation to the total observations in Fig. 2 Comparison of iSP-LDA, FP-LDA, HME-LDA, SLDA. Accuracy defines the identification of the Topic which is classified as appropriate from the listed data points.

$$\text{Accuracy} = TP+TN/TP+FP+FN+TN \text{ -------- (d)}$$

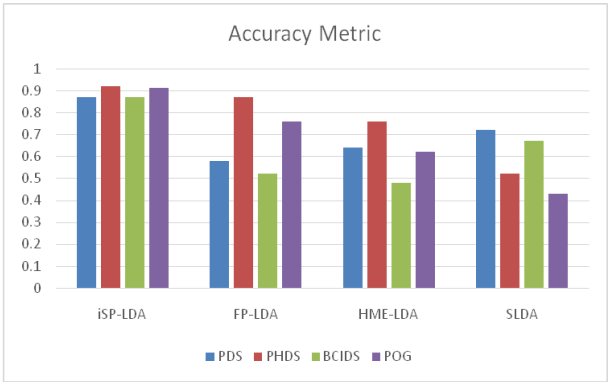


Fig.2. Comparison of iSP-LDA, FP-LDA, HME-LDA, SLDA

4.3 Precision:

Precision when defined for a classifier with a given class is the ratio of true positives which should be combined to both true and false positives. True Positives (TP) is the correctly predicted Topic for the given document in the data source which means that the value of actual class is yes, and the value of predicted class is also yes. False Positives (FP) – When actual class is no and predicted class is yes Fig. 3 Comparison of iSP-LDA, FP-LDA, HME-LDA, SLDA.

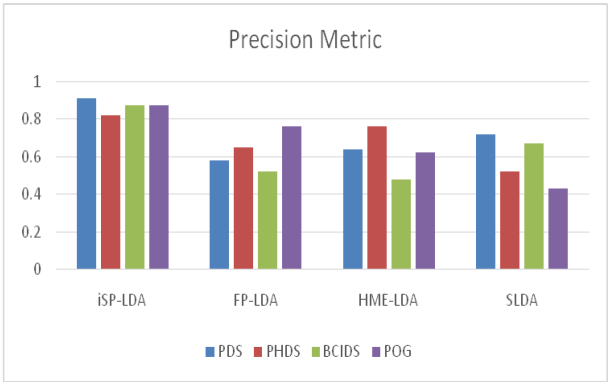


Fig.3. Comparison of iSP-LDA, FP-LDA, HME-LDA, SLDA

4.4 Recall:

Recall is calculated with the ratio of True Positive Spam Positive words identified from the data set through BCC Clustering which shares the base data source for the LIME classification technique Fig. 4 Comparison of iSP-LDA, FP-LDA, HME-LDA, SLDA.

$$Recall = TP/TP+FN \text{ ---------- (e)}$$

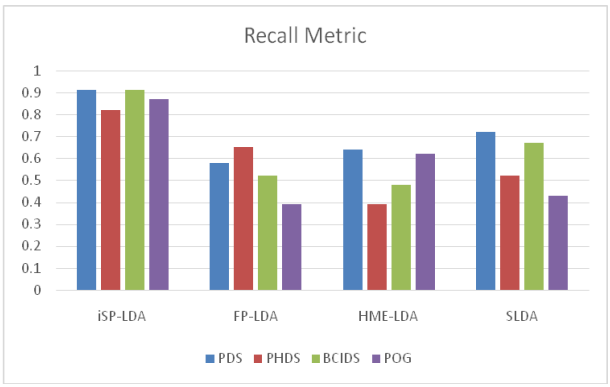


Fig.4. Comparison of iSP-LDA, FP-LDA, HME-LDA, SLDA

4.5 F1 Score:

The F1 score is calculated based on the weighted average of Precision and Recall. F1 Score = 2*(Recall * Precision) / (Recall + Precision), Based on this Graph and is plotted as below in Fig.5. Comparison of iSP-LDA, FP-LDA, HME-LDA, SLDA.

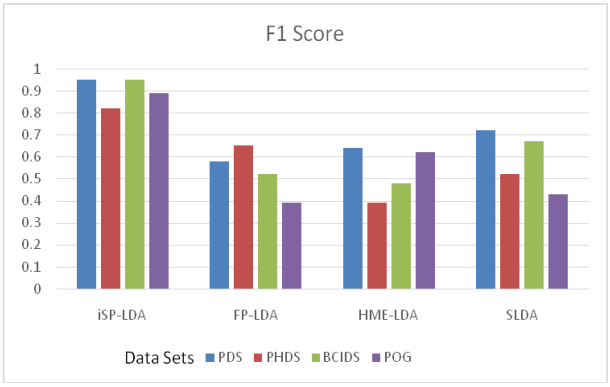


Fig.5. Comparison of iSP-LDA, FP-LDA, HME-LDA, SLDA

4.6 AUC:

Area under the curve (AUC) is calculated as the AUC of a classifier is equal to the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative.

P (score(x+)>score (x−))

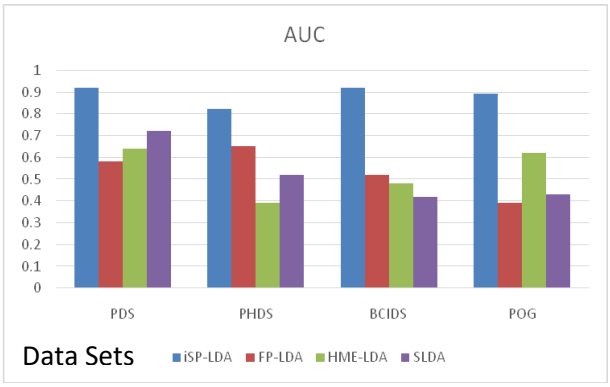


Fig.6. Comparison of iSP-LDA, FP-LDA, HME-LDA, SLDA

A set of statements or facts is said to be coherent, if they support each other. Thus, a coherent fact set can be interpreted in a context that covers all or most of the facts Fig.6 Comparison of iSP-LDA, FP-LDA, HME-LDA, SLDA. An example of a coherent fact set is “the game is a team sport”, “the game is played with a ball”, “and the game demands great physical efforts”.

4.7 Coherence Measures

Let’s take quick look at different coherence measures, and how they are calculated:

1. ***C_v*** measure is based on a sliding window, one-set segmentation of the top words and an indirect confirmation measure that uses normalized point wise mutual information (NPMI) and the cosine similarity
2. ***C_p*** is based on a sliding window, one-preceding segmentation of the top words and the confirmation measure of Fitelson’s coherence
3. ***C_uci*** measure is based on a sliding window and the point wise mutual information (PMI) of all word pairs of the given top words
4. ***C_umass*** is based on document co occurrence counts, a one-preceding segmentation and a logarithmic conditional probability as confirmation measure
5. ***C_npmi*** is an enhanced version of the C_uci coherence using the normalized point wise mutual information (NPMI)
6. ***C_a*** is based on a context window, a pair wise comparison of the top words and an indirect confirmation measure that uses normalized point wise mutual information (NPMI) and the cosine similarity

Topic Coherence measures score a single topic by measuring the degree of semantic similarity between high scoring words in the topic. These measurements help distinguish between topics that are semantically interpretable topics and topics that are artifacts of statistical inference Fig.7 Comparison of iSP-LDA, FP-LDA, HME-LDA, SLDA. It calculates how often two words, appear together in the corpus and it’s defined as

$$C_{UMass}(w_i, w_j) = \log \frac{D(w_i, w_j) + 1}{D(w_i)},$$

where $D(w_i, w_j)$ indicates how many times words w_i and w_j appear together in documents, and $D(w_i)$ is how many time word w_i appeared alone. The greater the number, the better is coherence score. Also, this measure isn't symmetric, which means that $C_{UMass}(w_i, w_j)$ is not equal to $C_{UMass}(w_j, w_i)$. We calculate the global coherence of the topic as the average pairwise coherence scores on the top N words which describe the topic.

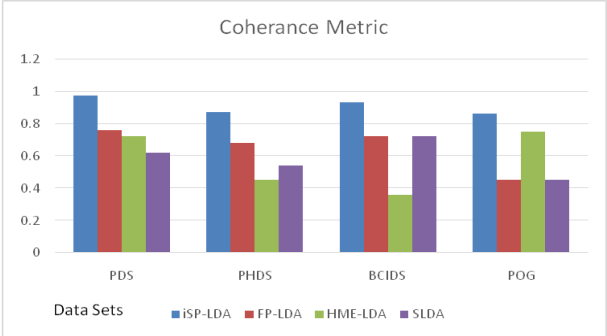


Fig.7. Comparison of iSP-LDA, FP-LDA, HME-LDA, SLDA

Discussion:

The datasets taken for the study also contained a similar set of articles across each unique topic. Vectorization techniques used for creating pattern dictionary and doc_term_matrix preserved the topic hierarchy of the documents. Also, it is to be noticed that iSPLDA methods provide the best learned descriptive topics compared to the other methods, aside from some methods that failed to create topics that aggregate related words. The iSPLDA model produces higher-quality topics and more coherent topics than the other methods in our evaluated datasets. Extracting fewer numbers of keywords led to a high coherence score in iSPLDA [10].

Conclusion:

In this paper, iSP-LDA a topic modeling based method for knowledge discovery is proposed. It works in unsupervised manner, gathering frequent semantic patterns using XLNet embedding on natural language data. Big data has information and knowledge sources that might confuse users and cause them to spend additional time and effort trying to find applicable information about specific topics or objects. Conversely, the need to analyze longer texts has become significantly relevant. The challenge with inferring topics from documents is due to the fact that it contains relatively larger amounts and noisy data that might result in inferring an inaccurate topic. iSPLDA can overcome such a problem since it incorporates semantic knowledge about domain as features to achieve topic categorization. The information regarding each topic obtained by the proposed model can be applied to numerous areas of study such as Information Retrieval, computational linguistics and NLP. Text pattern mining approaches still have challenges related to methods used to solve real-world tasks like scalability problems.

References:

[1] Aggarwal, A., & Toshniwal, D. (2019). Frequent pattern mining on time and location aware air quality data. IEEE Access, 7, 98921-98933.

[2] Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., ... & Warschauer, M. (2020). Mining big data in education: Affordances and challenges. Review of Research in Education, 44(1), 130-160.

[3] Ghojogh, B., Samad, M. N., Mashhadi, S. A., Kapoor, T., Ali, W., Karray, F., & Crowley, M. (2019). "Feature selection and feature extraction in pattern analysis: A literature review", arXiv preprint arXiv:1905.02845.

[4] Gurcan, F., &Cagiltay, N. E. (2019) "Big data software engineering: Analysis of knowledge domains and skill sets using LDA-based topic modeling", *IEEE Access*, *7*, 82541-82552.

[5] Jankowski, M. (2018, June), "Boost multi-class sLDA model for text classification". In International Conference on Artificial Intelligence and Soft Computing (pp. 633-644). Springer, Cham.

[6] Ko, A., & Gillani, S. (2020), "A research review and taxonomy development for decision support and business analytics using semantic text mining", International Journal of Information Technology & Decision Making, 19(01), 97-126.

[7] Kumar, S., Kar, A. K., & Ilavarasan, P. V. (2021), "Applications of text mining in services management: A systematic literature review", International Journal of Information Management Data Insights, 1(1), 100008.

[8] Li, Y., & Zhang, D. (2020, September), "Hadoop-Based University Ideological and Political Big Data Platform Design and Behavior Pattern Mining", In 2020 International Conference on Advance in Ambient Computing and Intelligence (ICAACI) (pp. 47-51), IEEE.

[9] Min, F., Zhang, Z. H., Zhai, W. J., & Shen, R. P. (2020) "Frequent pattern discovery with tri-partition alphabets", *Information Sciences*, *507*, 715-732.

[10] Saquicela, V., Baculima, F., Orellana, G., Piedra, N., Orellana, M., & Espinoza, M. (2018). Similarity Detection among Academic Contents through Semantic Technologies and Text Mining. In *IWSW* (pp. 1-12).

[11] Sundravadivelu, K, Senthilvel, P.G.,Duraimutharasan, N.,Esther T, H.R.,Kumar. K, R." Extensive Analysis of IoT Assisted Fake Currency Detection using Novel Learning Scheme", ICAISS 2023, pp. 1469–1477, 10.1109/ICAISS58487.2023.10250560.

[12] Sundravadivelu, K,Senthilvel, P.G.,Thirupurasundari, D.R.,Rajesh Kumar, K.,Palani, H.K., "Automated Drone-Based Imaging Systems For Plant Health Monitoring Using Deep LearningTechniques",2023, Intelligent Computing and Control for Engineering and Business Systems, ICCEBS 2023, IEEE , 10.1109/ICCEBS58601.2023.10449190.

[13] Sundravadivelu, K., Thangaraj, M., Gnanambal, S. (2022). An extensive work on comparing sentiment patterns in twitter archives between two persons. International Journal of Health Sciences,6(S7), 5170- 5180. https://doi.org/10.53730/ijhs.v6nS7.13104.

[14] Sundravadivelu. K, M. Thangaraj, "Analyzing Educational Tweets using LDA Model", International Journal of Intelligent Systems and Applications in Engineering (IJISAE), Volume 10, Issue 4, ISSN:2147-6799, PP. 100- 104, Dec. 2022.

[15] Sundravadivelu. K, M. Thangaraj, ―A Novel Approach for Discovering the Patterns by using PDBD Model in Big Data‖, Journal of Computer Science, (Science Publications), Volume 18 Issues 5, DOI: 10.3844 / jcssp.2022.382.395, ISSN: 1549-3636, pp.382-395, May 2022.

[16] Sundravadivelu. K, Suraj Rajesh Karpe, Harish V Mekali, Shital Nalgirkar, K. Abdul Rasak, Dr. V S Narayana Tinnalur, Information Theory and Coding: Techniques for Error Control and Data Compression, Journal of Electrical Systems 20-10s (2024): 5665-5674.

[17] Sundravadivelu. K, Thangaraj. M,"Mining effective patterns from text data - a survey", International Journal of Scientific and Technology Research, Volume 9, Issue 1, Pages 1930 – 1934, January 2020.

[18] Sung, S. F., Lin, C. Y., & Hu, Y. H. (2020). EMR-based phenotyping of ischemic stroke using supervised machine learning and text mining techniques. IEEE Journal of Biomedical and Health Informatics, 24(10), 2922-2931.

[19] Thangaraj.M.,&Sundaravadivelu.K, "Mining effective patterns from text data-a survey", International Journal of Scientific & Technology Research, ISSN-10: 2277-86161930 IJSTR, 2020.

[20] Wang, L. L., & Lo, K. (2021). Text mining approaches for dealing with the rapidly expanding literature on COVID-19. Briefings in Bioinformatics, 22(2), 781-799.[1]

[21] Xing, W., Lee, H. S., &Shibani, A. (2020). Identifying patterns in students' scientific argumentation: content analysis through text mining using Latent Dirichlet Allocation. *Educational Technology Research and Development, 68*(5), 2185-2214.

[22] Yiran, Y., & Srivastava, S. (2019, June). Aspect-based Sentiment Analysis on mobile phone reviews with LDA. In *Proceedings of the 2019 4th International Conference on Machine Learning Technologies* (pp. 101-105).

[23] Yuan, L., Bin, J., Wei, Y., Huang, F., Hu, X., & Tan, M. (2020) "Big data aspect-based opinion mining using the slda and hme-lda models. *Wireless Communications and Mobile Computing", 2020*.

[24] Zulkefli, N. S. S. B., Rahman, N. B. A., Puteh, M. B., & Bakar, Z. B. A. (2018, March). "Effectiveness of Latent Dirichlet allocation model for semantic information retrieval on Malay document", In *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)* (pp. 1-5), IEEE.