JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Latest Trends in Latent Semantic Analysis Method and Comprehensive Preview

[1]Taralkumar Contractor, [2]Mr. Jitendrakumar Dhobi, [3]Mr. Nitin Raval
[1]Post Graduate Student, [2]Professor, [3]Assistant Professor
Department of Computer Engineering,
Government Engineering College, Sector-28, Gandhinagar

Abstract
Latent Semantic Analysis (LSA) is a potent tool aimed at bridging the gap between human language comprehension and machine understanding. Recognizing that language extends beyond mere words and faces challenges of high dimensionality in textual data, LSA operates on the premise that documents and words sharing similar contexts likely have semantic connections. This paper delves into the foundational principles of LSA, its applications, and its impact across various domains. LSA is a fully automated mathematical approach that extracts and infers relationships among words' expected contextual usage, utilizing techniques such as Document Term Matrix and Singular Value Decomposition. It serves as a versatile method for comprehending and learning from text, providing valuable insights to enhance information retrieval and filtering. As the field of Natural Language Processing progresses, LSA remains a fundamental technique for extracting insights from textual data.

Index Terms - Latent Semantic Analysis, Document Term Metrix, Singular Value Decomposition

I. INTRODUCTION

In today's Digital world, everything has been shifted to the online. There is a gigantic amount of data available on the internet. The data is available in every field: education, business, politics, health, history, legal, commerce, sports etc. The information retrieval is challenging role for Artificial Intelligence by analyzing massive volumes of data and recognizing patterns in the data given. Natural Language Processing is a branch of AI that focuses on helping machines to understand human language and build systems that can make sense of text and automatically perform tasks. Natural Language Processing working in 5 steps as below mention:

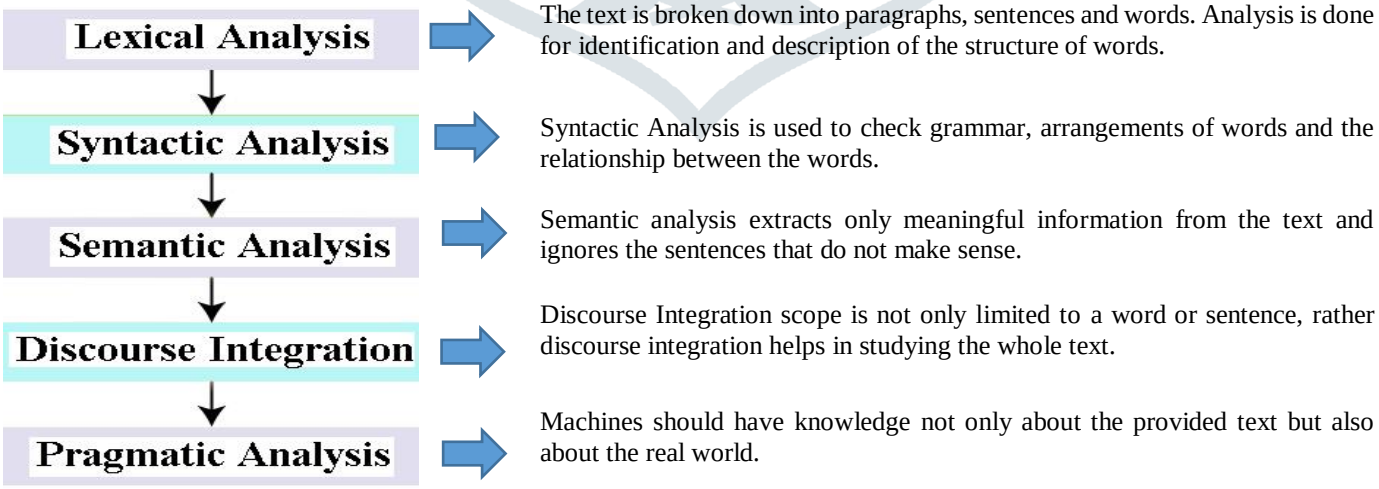

Figure 1.1: Natural Language Processing steps

Semantic Analysis attempts to understand the meaning of Natural Language. Semantic Analysis of Natural Language captures the meaning of the given text while considering context, logical structuring of sentences, and grammar roles. Semantic analysis can begin with the relationship between individual words. Semantic Analysis are divided into two parts.

1.Latent Semantic Analysis
2.Compositional Semantics Analysis.

Latent Semantic Analysis

Latent Semantic Analysis is mathematical technique for retrieving and revealing hidden relations of contextual usage of words in document.

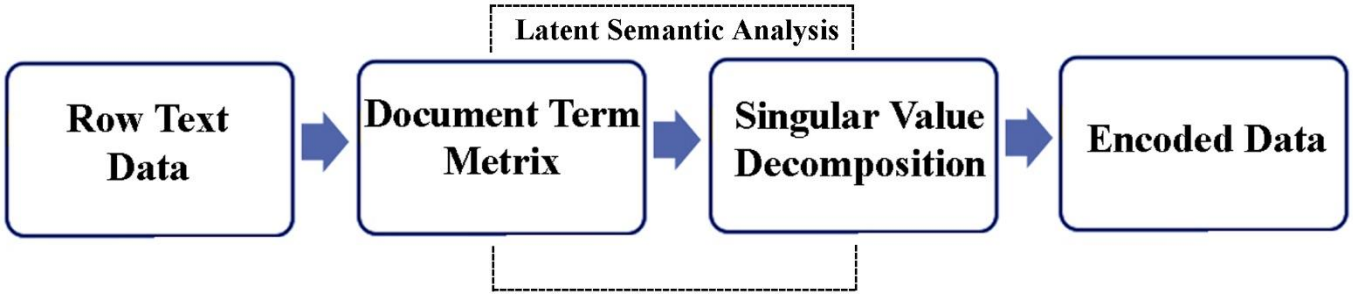

Fig.1.2: LSA Basic Flow Diagram

Document Term Matrix

A Document-Term Matrix is like a numerical map of a collection of texts. Each row stands for a document, while each column represents a specific word found in all the documents put together. The numbers in the matrix show how often each word appears in each document.

Singular Value Decomposition

Singular Value Decomposition (SVD) is a key mathematical method used in many fields, including Latent Semantic Analysis (LSA). It breaks down a matrix into three separate matrices, offering a method to represent the original matrix in a more concise and insightful manner.

For a given matrix A (of size m x n), SVD decomposes it into three matrices:

$\mathbf{A = U\Sigma V^T}$

U: The matrix U, consisting of m rows and m columns, contains orthogonal vectors known as the left singular vectors. These vectors create a set of mutually perpendicular directions that serve as a basis for the columns of the original matrix A. They effectively represent the interrelationships among the rows in the original matrix.

Σ : The matrix Σ, with dimensions m by n, is a diagonal matrix that holds the singular values of matrix A. These singular values are always positive and signify the square root of the eigenvalues of either ATA or AAT. They are arranged in decreasing order along the diagonal of Σ.

V^T : The matrix V^T, with dimensions n by n, comprises orthogonal vectors known as the right singular vectors. These vectors establish a set of mutually perpendicular directions that serve as a basis for the rows of the original matrix A. They effectively represent the interrelationships among the columns in the original matrix.

2. TRENDS IN LATENT SEMANTIC ANALYSIS

1. Text Summarization for News Articles using Latent Semantic Analysis Technique

Text summarization has been a longstanding subject in academic circles, with various automatic techniques emerging in recent years. However, achieving efficiency remains a challenge. With the vast amount of online documents, there's a growing need for automatic news summarization. This study proposes a text summarization approach that emphasizes identifying crucial text segments and generating coherent summaries without delving deeply into text semantics. Instead, it utilizes Latent Semantic Analysis to construct summaries efficiently.

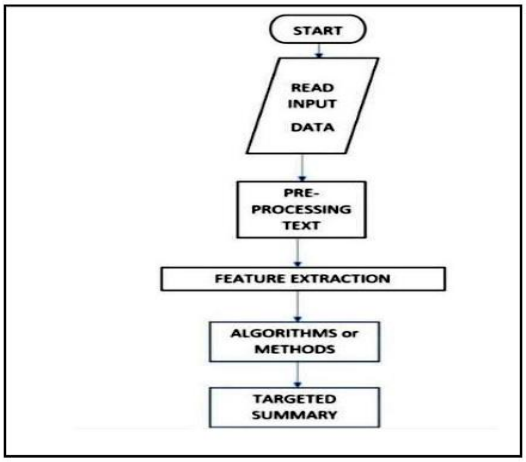

Figure 2.1: Block Diagram for Proposed Method [1]

Latent Semantic Analysis (LSA) is a robust statistical technique that uncovers implicit features within words and sentences, revealing their underlying semantic structures. Combining Natural Language Processing with unsupervised methods, it's employed in text summarization projects to distill key information from text. The process involves preprocessing the text by cleaning and tokenizing it, then constructing a term-document matrix and applying Singular Value Decomposition (SVD) to reduce its dimensions. Cosine similarity is utilized to gauge similarity between the summary and documents in the reduced space. Finally, documents with the highest similarity scores are chosen as the most pertinent, forming the foundation of the text summary.

2. Adverse Drug Reaction Detection Using Latent Semantic Analysis

Detecting Adverse Drug Reactions (ADRs) is crucial for understanding patients' experiences with medications. ADRs refer to unintended and harmful reactions to medications, which can range from mild discomfort to severe health complications. Monitoring and identifying ADRs allow healthcare providers to assess the safety and effectiveness of drugs and make informed decisions about patient care. This study proposes a semantic method based on Latent Semantic Analysis (LSA) to enhance the detection of Adverse Drug Reactions (ADRs). Traditionally, detection methods relying solely on trigger terms may lack the ability to capture latent semantic associations between words and phrases, potentially leading to missed ADR occurrences. To address this limitation, the study utilizes LSA, a technique that analyzes relationships between terms and documents in a corpus to uncover latent semantic similarities. The proposed method is applied to a benchmark dataset, with several preprocessing steps implemented including stop word removal, tokenization, and stemming. n addition to the semantic method based on Latent Semantic Analysis (LSA), this study employed two representations of documents: Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF).

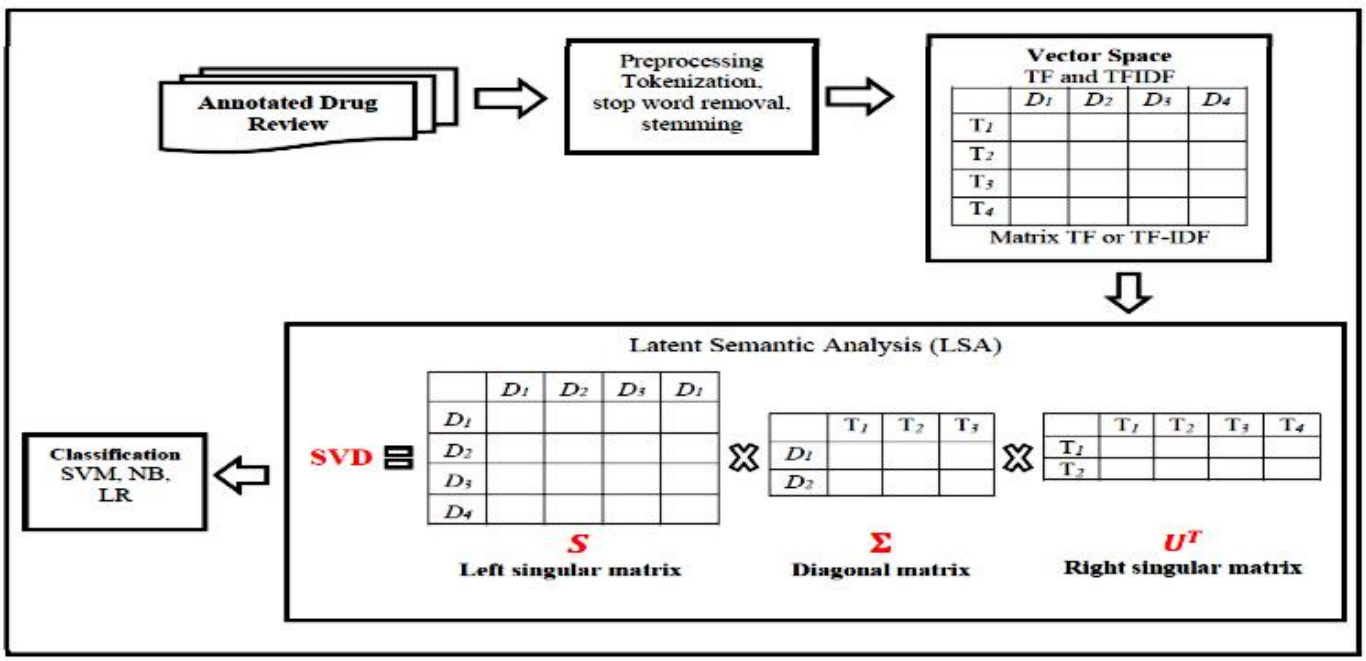

Figure 2.2: Proposed LSA Methods[2]

Results indicated that the proposed LSA approach surpassed the baseline method of using extended trigger terms. LSA achieved an F-measure of 82% for the dataset, showcasing its superiority over trigger term-based approaches. This performance improvement underscores the effectiveness of LSA in identifying semantic correspondences accurately, as opposed to relying on a predefined list of trigger terms. In summary, the study demonstrates that incorporating LSA for ADR detection outperforms traditional methods based on trigger terms. By leveraging semantic analysis, LSA enables more precise identification of ADRs, resulting in improved accuracy and effectiveness in detecting adverse drug reactions[(2)].

3. Sentence Similarity Using Modified Latent Semantic Analysis and Semantic Relations

This paper introduces a method for determining the similarity between pairs of sentences by combining two modules: a modified Latent Semantic Analysis (LSA) and semantic similarity computation. The method leverages both the syntactic structure and semantic information inherent in the sentence pairs. First, the syntactic structure, represented as dependency triplets, is extracted from the sentence pairs. Then, semantic similarity between words is calculated using the Wu & Palmer similarity measure and WordNet synonym relations are incorporated into the modified LSA.

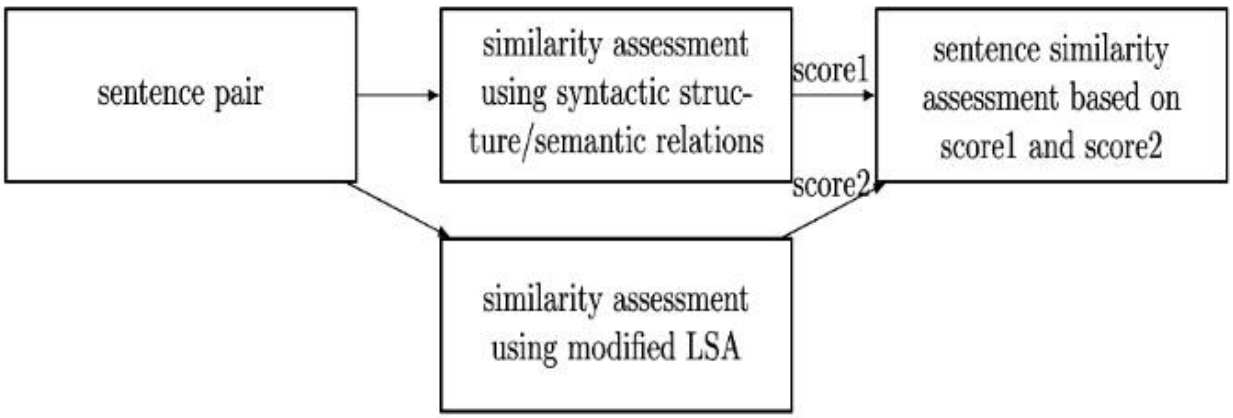

Figure 2.3: Block diagram of Proposed Method [(3)]

The proposed method is evaluated on two datasets: the Microsoft Research Paraphrase Corpus and the Li et al. text similarity dataset. For the Microsoft Research Paraphrase Corpus, the method achieves an accuracy of 73.19%, outperforming existing statistical and zero-shot domain adaptation methods. Additionally, on the Li et al. text similarity dataset, the method achieves high correlation coefficients of 0.9021 (Pearson) and 0.9103 (Spearman), along with a mean deviation of 0.105 compared to human judgment, demonstrating its superiority over state-of-the-art methods. In summary, the proposed method combines syntactic and semantic features to accurately measure sentence similarity, surpassing existing approaches on two benchmark datasets [(3)].

4. A New Methodology for Computing Semantic Relatedness: Modified Latent Semantic Analysis by Fuzzy Formal Concept Analysis

With the exponential growth of electronic information, efficiently comprehending vast amounts of text has become a significant concern in research communities. The sheer volume of textual information on the web, estimated to be around 50 million pages, surpasses what humans can interpret manually. Therefore, there's a necessity for computer understanding to handle this immense textual data. To interpret such large text datasets effectively, it's crucial to understand the meanings of different words. This understanding involves more than just recognizing individual words; it requires grasping their contextual nuances, semantic relationships, and connotations. Through techniques such as natural language processing (NLP), machine learning, and semantic analysis, computers can learn to comprehend the meaning of words in various contexts and extract valuable insights from vast amounts of textual data. This enables automated processes for tasks like information retrieval, sentiment analysis, summarization, and more, thereby facilitating efficient utilization of textual information on the web. In the literature, two main methods for measuring semantic similarity between words are distinguished: Corpus-based methods and Knowledge-based methods. Corpus-based methods rely on utilizing large amounts of text data (corpora) to infer semantic relationships between words. These methods utilize statistical and co-occurrence-based techniques to calculate the association among words. If two words frequently appear together in the same context within the text data, they are considered more similar or associated. These measures primarily focus on analyzing the co-occurrence patterns of words within the corpus. Common techniques used in corpus-based methods include distributional semantics, word embedding and probabilistic models like Latent Semantic Analysis (LSA) and Word2Vec.

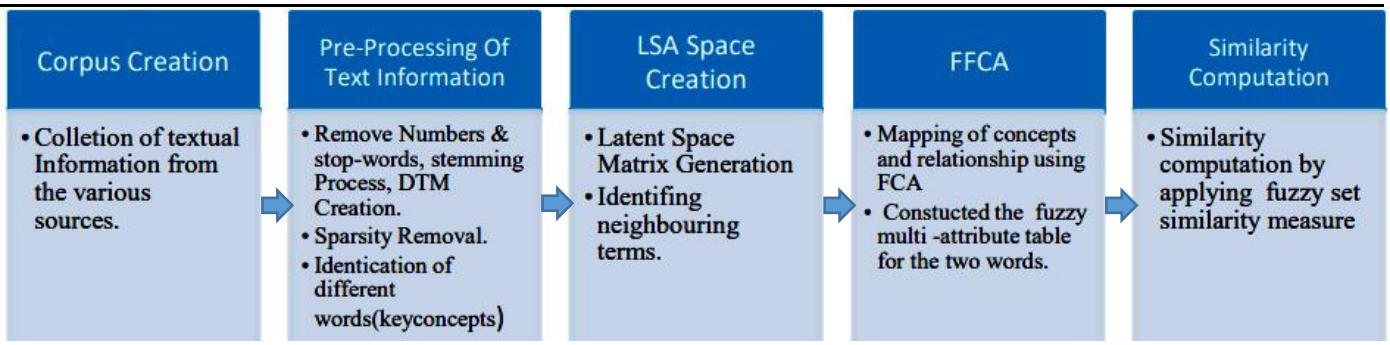

Fig.2.4: Process Flow Diagram [4]

In this research, a hybrid approach of LSA, FFCA (Fuzzy formal concept Analysis) and fuzzy similarity measure is proposed for finding semantic relatedness among words. Fig 1 shows the process flow diagram of the proposed methodology. In this paper, a novel methodology is introduced to determine semantic relatedness among words, combining Latent Semantic Analysis (LSA) and Fuzzy Formal Concept Analysis (FFCA). Initially, LSA is employed, where similarity is calculated based on word frequency relative to the total frequency of words. However, LSA's accuracy is deemed lower compared to knowledge-based approaches. To enhance accuracy, neighboring terms obtained from LSA are utilized as features or attributes for concepts. Subsequently, similarity is computed based on common and unrelated words. In summary, the paper introduces a methodology that combines LSA and FFCA to compute semantic relatedness among words, enhancing accuracy by incorporating neighboring terms and utilizing a fuzzy set similarity measure. The application of this method in the solar domain demonstrates its effectiveness in achieving improved accuracy compared to existing approaches [4].

5. Using Latent Semantic Analysis to Identify Research Trends in OpenStreetMap

In this paper, Latent Semantic Analysis (LSA) is employed to analyze a corpus consisting of 485 academic abstracts related to OpenStreetMap (OSM) research published between 2007 and 2016. The objective is to identify emerging research trends within the OSM academic community. Using LSA, the study identifies five core research areas and fifty distinct research trends within the OSM domain. By analyzing the semantic relationships between words and concepts in the abstracts, LSA helps uncover underlying patterns and themes across the literature. The findings of the study provide valuable insights into the evolving landscape of OSM research, shedding light on the key topics, themes, and areas of focus within the academic community during the specified time frame. In this paper Data Acquisition, Application of Latent Semantic Analysis, Pre-Processing and Term-Filtering, Term Frequency-Inverse Document Frequency, Singular Vector Decomposition, Dimensional Reduction: Selecting Optimal Topic Solutions, Selecting Threshold Values for Topic Solutions, Topic Labeling techniques used for proper retrieval of information. In the domain of OpenStreetMap (OSM) research, the "quality assessment and analysis" area has received significant attention from researchers, particularly focusing on specific regions of the world. Scholars have dedicated efforts to identifying intrinsic quality indicators, recognizing that established measures alone are insufficient for evaluating the quality of OSM data. Additionally, another prominent research area that has emerged is centered around understanding the motivations and patterns of contributors, falling under the category of "assessment of contributors' behavior." Researchers are keen on comprehending the driving forces behind contributors' engagement with OSM, as well as their behaviors and patterns of participation. In summary, within the realm of OSM research, considerable emphasis has been placed on assessing data quality and understanding contributors' behavior. These areas represent critical avenues of investigation, aiming to enhance the reliability and effectiveness of OSM data while delving into the motivations and dynamics of the community contributing to its development [5].

6. Three level weight for latent semantic analysis: an efficient approach to find enhanced semantic themes

In this paper, the authors propose an enhanced approach for Latent Semantic Analysis (LSA) aimed at optimizing semantic space for large collections of documents. This three-level weighting scheme aims to bring terms in documents closer together, even if they appear distant in the original document collection. The approach is evaluated on both synthetic datasets comprising small stories and the BBC-news dataset commonly used in text mining applications.

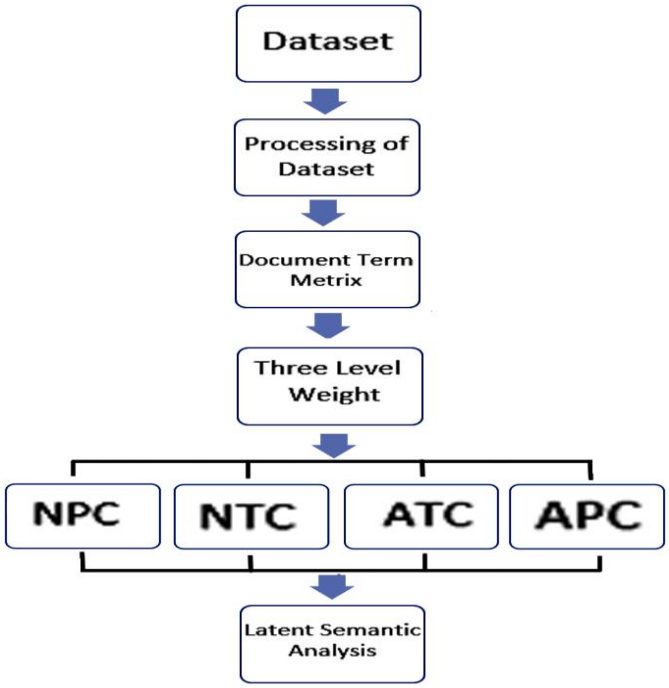

Figure 3.1: Methodology for Three Level Latent Semantic Analysis [6]

The proposed three-level weight models, namely NPC, NTC, APC, and ATC, assign weights at the term, document, and corpus levels. These models are tested against traditional term frequency methods and exhibit notable improvements in term set correlation, document set correlation, and semantic similarity of terms in the generated semantic space. Moreover, the authors demonstrate the automatic context clustering capabilities of their approach, showcasing its effectiveness in organizing datasets based on semantic relationships. Overall, the proposed methodology presents a promising advancement in semantic themes detection and topic modelling, offering improved performance over existing techniques across various evaluation metrics and datasets [6]. In addition to improving semantic analysis, the implementation of the three-level weighted model for LSA also reduces computational complexity. By incorporating these weighted models, the algorithm can streamline the processing required for semantic analysis, making it more efficient and scalable.

In this experiment, four different versions of three levels of weighted models for Latent Semantic Analysis (LSA) were implemented: NTC NPC, ATC, and APC. Cosine normalization of document length was applied in all four versions. The objective was to evaluate the impact of these weighted models on LSA by analyzing the semantic space generated through the Singular Value Decomposition (SVD) process, which produces three matrices: term matrix, document matrix, and singular value matrix.

To analyze term correlation, cosine similarity scores between vectors in the term matrix (∑U matrix) were calculated. For document correlation analysis, cosine similarity between vectors in the document matrix (∑VT matrix) was calculated. This relatedness information facilitated the organization of terms and documents into semantic groups, where terms and documents sharing the same semantic space were clustered together, forming topics.

LSA enhances the grouping of documents into clusters of interest by achieving high correlations between documents. The semantic space generated three matrices: term matrix Tk, document matrix Dk, and singular matrix Sk. This process enables the clustering of related terms and documents into topics, thereby improving the organization and understanding of the underlying semantic structure within the document corpus [6].

7. Practical use of a latent semantic analysis (LSA) model for automatic evaluation of written answers

Research on automatic evaluation systems for written texts has been ongoing since the 1960s, aiming to analyze and rate written responses, particularly in educational contexts to measure students' learning degree. However, it wasn't until the late 1990s that advancements in natural language processing (NLP) led to the development of new models and methods, resulting in higher accuracy levels suitable for practical applications. Recent studies indicate that automatic evaluation systems have achieved accuracy levels closer to that of human evaluators, marking significant progress in the field. This progress suggests the increasing potential of computational technologies to effectively assess written texts, offering valuable insights for various educational and evaluative purposes.

This paper presents research on applying a latent semantic analysis (LSA) model for automatically evaluating short answers (25 to 70 words) to open-ended questions. The research aims to enhance the robustness, accuracy, and portability of the LSA model. To

achieve this, the methods included implementing word bigrams, combining unigrams and bigrams using multiple linear regression, and adding an adjustment step after score attribution based on the average word count of the answers.

The study utilized a corpus of 359 answers from a Brazilian public university's entrance examination, previously scored by human evaluators. The experiments resulted in an accuracy of approximately 84.94%, while human evaluators achieved an accuracy of about 84.93%. These results suggest that the automatic evaluation technology is approaching a high level of efficiency, demonstrating the potential for reliable automated assessment of short answer responses to open-ended questions.

8. Latent Semantic Analysis Based Sentimental Analysis of Tweets in Social Media for the Classification of Cyberbullying Text

Recently, the importance of technologies capable of automatically detecting potential cyberbullying behaviors has become evident, as they can aid in preventing harmful situations for victims. Despite the increasing societal concern surrounding cyberbullying, there has been relatively little computer research on the topic thus far. The development of such technologies holds promise for addressing cyberbullying effectively by automatically identifying behaviors associated with it. This underscores the need for further research and development in this area to mitigate the negative impacts of cyberbullying.

With the widespread use of mobile technology, cyberbullying has become a significant issue, particularly among adolescents, leading to tragic outcomes such as suicide. As a result, there is a growing awareness of the problem within society. Despite various efforts to address cyberbullying, many existing methods for detecting cyberbullying texts lack accuracy. This paper proposes a Latent Semantic Analysis (LSA) based sentiment analysis approach for classifying cyberbullying text in social media. By employing LSA, the study aims to accurately classify texts, thereby enabling users to express their opinions on social media platforms without experiencing online abuse. Simulation results demonstrate that the proposed method achieves a higher accuracy rate compared to existing methods, highlighting its potential effectiveness in combating cyberbullying.

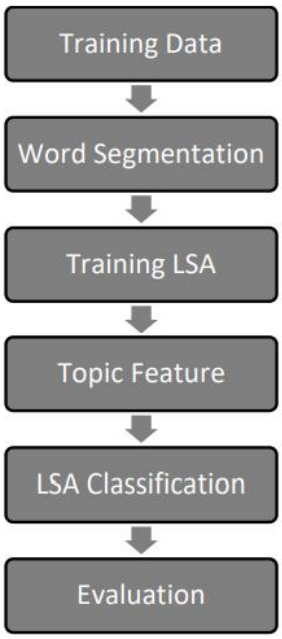

Figure 3.1: Proposed Method[8]

The research introduces a novel LSA-based model for cyberbullying detection, employing sentiment analysis as its core method. In this model, cyberbullying is considered inherently negative, simplifying the classification task. To avoid the laborious process of manually labeling large datasets and making prior assumptions about class distributions, the research opts for an unsupervised technique. This approach streamlines the process and saves considerable time by leveraging LSA's ability to uncover latent semantic patterns in the data without explicit class labels.

3. COMPARATIVE STUDY

Table 3.1: Discussion of comparative studies

Serial Number	Title	Technique Used	Point to Consider	Result
1.	Text Summarization for News Articles using Latent Semantic Analysis Technique	Preprocessing, LSA, Similarity Calculation, Ranking and Selection	Preprocessing, Stop words removal, LSA, Text Summarization	Text Summarization done by LSA for news articles.
2	Adverse Drug Reaction Detection Using Latent Semantic Analysis	Preprocessing (Stop word removal, Tokenization, Stemming), Term Representation (Tf-IDF) , SVD, Classification (SVM, NB, LR)	SVD, Classification helps to improve Performance	The Proposed LSA get better results by TF feature and LR classifier was used for detecting ADRs..
3	Sentence Similarity Using Modified Latent Semantic Analysis and Semantic Relations	Modified LSA, Similarity assessment using syntactic structure / Semantic relations, Cosine similarity	Sentence similarity assessment based on score1 and score2	Modified LSA is better than statistical , Zero-shot domain adaptations and state-of-the-art methods.
4.	A New Methodology for Computing Semantic Relatedness: Modified Latent Semantic Analysis by Fuzzy Formal Concept Analysis	Corpus Creation and Pre-processing, LSA Space Creation, Fuzzy Formal Concept Analysis (FFCA)	FFCA and fuzzy set similarity measure,	LSA and fuzzy formal concept analysis is used to compute the semantic relatedness.
5.	Using Latent Semantic Analysis to Identify Research Trends in OpenStreet Map	DTM, Pre-Processing and Term-Filtering, TF-IDF, SVD, Dimensional Reduction, Selecting Threshold Values for Topic Solution, Topic Labelling	GIS(Geographic Information System), LSA, OSM(OpenStreetMap), SVD, TF-IDF,VGI (Volunteered Geographical Information)	LSA is very useful to Identify Research trends in OpenStreetMapand also helped general recommendations regarding research gaps that have emerged from the different core research area.
6.	Three level weight for latent semantic analysis: an efficient approach to find enhanced semantic themes	Preprocessing of data (like stop-word, stemming), Term frequency, Collection frequency, Document length, SCD, Cosine Similarity TF,Probability based IDF, Augmented TF, TFbased IDF	Three level – Term level, Document Level, Document length Normalization, LSA NTC, NPC, ATC and APC	In this research three level weights models in LSA which improved results in multiple dimensions in term and DTM generated by LSA.

7.	Practical use of a latent semantic analysis (LSA) model for automatic evaluation of written answers	Preprocessing, Weighing, SVD, Rating, Adjustments, Accuracy	Co-occurrence of unigrams and bigrams, Combine bigrams and LSA	The results improved by from methods based on n-grams. Accuracy 84.93%.
8.	Latent Semantic Analysis Based Sentimental Analysis of Tweets in Social Media for the Classification of Cyberbullying Text	Preprocessing, LSA Classification, KNN, NB,	DTM SVD	LSA model for the detection of cyberbullying worked well by save a significant amount of time.

4. CONCLUSION

LSA has been pivotal in advancing NLP, uncovering hidden semantic connections in text. It has found success in various applications such as Text Summarization, Drug Reaction Detection, Sentence Similarity Checking, and more. LSA's strengths lie in revealing hidden topics and improving information retrieval.

REFERENCES

[1] Rajalakshmi R, Vidhya S, Harina D, Karna R, Sowmya A. Text Summarization for News Articles using Latent Semantic Analysis Technique. In2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC) 2023 Jul 6 (pp. 1421-1425). IEEE.

[2] Nafea AA, Omar N, AL-Ani MM. Adverse drug reaction detection using latent semantic analysis. Journal of Computer Science. 2021 Oct;17(10):960-70.

[3] Learning M, Kim JL. Sentence Similarity Using Modified Latent Semantic Analysis and Semantic Relations. Machine Learning and Artificial Intelligence: Proceedings of MLIS 2023. 2023 Nov 9;374:43.

[4]NishyReshmi S, Shreelakshmi R. Sentence Similarity Using Modified Latent Semantic Analysis and Semantic Relations. Machine learning and artificial intelligence. 2023;43;51

[5] Jain S, Seeja KR, Jindal R. A new methodology for computing semantic relatedness: Modified latent semantic analysis by fuzzy formal concept analysis. Procedia Computer Science. 2020 Jan 1;167:1102-9.

[7] Sehra SS, Singh J, Rai HS. Using latent semantic analysis to identify research trends in openstreetmap. ISPRS International Journal of Geo-Information. 2017 Jul 1;6(7):195.

[6] Kherwa P, Bansal P. Three level weight for latent semantic analysis: an efficient approach to find enhanced semantic themes. International Journal of Knowledge and Learning. 2023;16(1):56-72.

[7] Alves dos Santos JC, Favero EL. Practical use of a latent semantic analysis (LSA) model for automatic evaluation of written answers. Journal of the Brazilian Computer Society. 2015 Dec;21(1):21.

[8] Wise DJ, Ambareesh S, Ramesh Babu P, Sugumar D, Bhimavarapu JP, Kumar AS. Latent Semantic Analysis Based Sentimental Analysis of Tweets in Social Media for the Classification of Cyberbullying Text. International Journal of Intelligent Systems and Applications in Engineering. 2024;12(7s):26-35.