The Scientific Temper (2025) Vol. 16 (8): 4689-4703
E-ISSN: 2231-6396, ISSN: 0976-8653
Doi: 10.58414/SCIENTIFICTEMPER.2025.16.8.14
https://scientifictemper.com/

RESEARCH ARTICLE

Lancaster sliced regressive keyword extraction based semantic analytics on social media documents

Sharada C[1*], T N Ravi [2], S Panneer Arokiaraj[3]

Abstract

Semantic analytics is one of the new issues materialized in natural language processing (NLP) with the emergence of social networks. Semantic analytics on social media documents refers to the procedure of employing NLP techniques for analyzing the deeper sense and context of text on social media platforms. Making use of the amount of information being now available, research and industry have attempted materials and mechanisms to analyze sentiments automatically in social networks. It just goes beyond keyword exploration to understand the associations between words, phrases and concepts within a social media post, recognizing for a more refined clarification of user sentiment and purpose. While the extensive greater part of these days researchare is completely concentrating on enhancing the algorithms employed for sentiment evaluation, the present one emphasizes the advantages of employing a semantic-based method for representing the analysis' results, the emotions and social media specific concepts. In this work a method called, Lancaster Tokenized Sliced Inverse Regressive Keyword Extraction (LT-SIRKE) for performing efficient semantic analysis on social media documents is introduced. LT-SIRKE technique is divide as query pre-processing as well as keyword extraction. Initially in LT-SIRKE method, the user inputs their query into the user window. Afterward, the query is sent to the system for efficient pre-processing. In query pre-processing phase, Stochastic Gradient Descent Keras-based tokenization, Lancaster-based stemming and Zipf's Law-based stop word removal process is carried out. After preprocessing, keywords are extracted using Bayesian Averaging and Sliced Inverse Regression-based Keyword Extraction to facilitate efficient information access. Experimental assessment is performed with various metrics namely precision, recall, accuracy, keyword extraction time and error with number of user requested queries.

Keywords: Semantic Analytics, Natural Language Processing, Social Media, Lancaster Tokenized, Sliced Inverse Regression, Keyword Extraction.

关键词： 语义分析、自然语言处理、社交媒体、兰开斯特标记、切片逆回归、关键词提取。

[1]Research Scholar, Department of Computer Science, Thanthai Periyar Government Arts and Science College, Affiliated to Bharathidasan University, Tiruchirappalli, Tamil Nadu, India.

[2]Associate Professor, Department of Computer Science, Jamal Mohamed College, Affiliated to Bharathidasan University, Tiruchirappalli, Tamil Nadu, India.

[3]Associate Professor, Department of Computer Science, Thanthai Periyar Government Arts and Science College, Affiliated to Bharathidasan University, Tiruchirappalli, Tamil Nadu, India.

***Corresponding Author:** Sharada C, Research Scholar, Department of Computer Science, Thanthai Periyar Government Arts and Science College, Affiliated to Bharathidasan University, Tiruchirappalli, Tamil Nadu, India., E-Mail: csharada@gmail.com

How to cite this article: Sharada, C., Ravi, T.N., Arokiaraj, S.P. (2025). Lancaster Sliced Regressive Keyword Extraction Based Semantic Analytics on Social Media Documents. The Scientific Temper, **16**(8):4689-4703.

Doi: 10.58414/SCIENTIFICTEMPER.2025.16.8.14

Source of support: Nil

Conflict of interest: None.

Introduction

The procedure of categorizing manuscript documents into various categories is determined to text classification (Atandoh *et al*., 2023). It is measured as the final process of text categorization. The most present day developments in DL for text classification concentrate on increasingly complicated methods based on neural networks to handle huge datasets (Atandoh *et al*., 2023).

Two-stage deep learning pipeline was performed into remove cause-and-effect trio by employing causal graph database (Hershowitz *et al*., 2023). The method of sentence-level noise removal was utilized to eliminate immaterial data to causal semantics. In second stage, the combined joint entity-and-relation extraction model to eradicate the causal relations. On the way, to carry out the noise elimination and causality removal an annotated dataset of 1027 WRN records were determined. But, the different metrics does not incorporate hierarchical representations of failure modes.

A BERT included with deep learning known as Multi Layered Convolutional Neural Network (B-MLCNN) was

Received: 28/07/2025 Accepted: 03/08/2025 Published: 30/08/2025

proposed with intent of classifying sentiments as in textual reviews of manuscript (Atandoh *et al*., 2023). Here, every review was measured as a part document. Also, the method possessed the advantages of equally BERT and multi-layer CNN. The precision, recall and accuracy were improved by designed method.

An automated method called Latent Dirichl*et al*location (LDA) was designed with intend of facilitating detection and analysis of reports via environmental content (Qiu *et al*., 2024). Based on their geological content, the method employed existing NLP advantage along with and text mining to enable swift scanning. Additionally, an auto completion-driven report was generated with accurate results.

The BERT and robust NLP model uses deep neural network architecture based on transformer model. The BERT model is varied from conventional NLP models which course text one word at a moment. On the other hand, transformers process entire text input at a time that aids in acquiring the correlations between words more efficiently. However, BERT cannot lay hold of fundamental reasoning or acquire knowledge not explicitly stated.

With aim of the searching and extracting data from massive electronic texts through LSTM neural network model is classify the text information. However, they have disadvantages including computational complexity and sensitivity to data quality and quantity.

LSTM neural network was introduced for document classification by using distribution network planning (Yishun *et al*., 2023). The designed neural network performed data preparation and training process. The technical aspect of neural network here remained in efficiently processing large quantity of data and to provide high accuracy while performing classification. However, the classification accuracy was not improved by LSTM neural network.

A new document embedding method was introduced for clustering employing graph auto encoder (Jung 2023). An undirected and weighted sparse graph was built from numerous documents. Every text was specified by node. All weighted edges generated in graph have higher cosine similarities among end nodes. The graph auto encoder was utilized to obtain the node embedding vectors. All nodes embedding vector in graph was determined as document embedding vector. But, keyword extraction process failed to perform by document embedding method.

More than the current years, OSNs have materialized in vogue for distribution various types of data and shaping public opinion. But, this convenience has also caused to extensive diffusion of false news. The lack of rule has resulted in propagation of minimum quality and fake content, affectation an important threat as a whole. A novel semantic deep learning technique was applied to classify documents with the intent of identifying fake news accurately (Alghamdia *et al*., 2024). Yet another integrated methodology was proposed employing artificial neural networks to design moderate actions on social media contents (Galamiton *et al*., 2024).

A systematic literature review on document-based sentiment analysis employing deep learning was designed (Alshuwaier *et al*., 2022). Given the favorable results acquired by deep-learning algorithms, the interpretability of predictions has become notable as far as practical applications are concerned. A method was presented to generate semantic and quantitative explanations both accurately and precisely (Xia *et al*., 2021). A holistic survey on text classification employing deep learning techniques was investigated (Li *et al*., 2022).

With the swift evolution of information technology, online information has been increasing gradually in an exponential fashion, specifically in the form of text documents. Owing to this, text mining has received a considerable amount of interest. A systematic review employing BERTopic modeling was presented (Nedungadi *et al*., 2025). State-of-the-art NLP methods were briefed (Choi *et al*., 2022).

A novel hybrid sentiment analysis by classifying documents employing convolutional neural network and hidden markov was designed with improved performance (Najafabadi 2024). Numerous sources are present for knowledge extraction, however, till now, unstructured text is considered as the largest available knowledge source. These data in digital format need to manage. A plethora of research is done in this domain and several classifiers have been evolved. Ensemble of deep learning was proposed for classifying text document therefore increasing classifier accuracy (Ranjan *et al*., 2021). Yet another method to handle lengthened keywords was presented (Kukkar *et al*., 2023).

Literature Review

Document classification and semantic examination are two of the most typical NLP tasks with many rising applications used in numerous domains, to name a few being, health care and procedure making. Together the huge growth in popularity and custom of social media has resulted on a massive up surge in user-driven data. But, the analysis of these detailed data in documents is a difficult task due to field variety to describe these data. Social media word-of-mouth analysis integrating AViT model and EBERT model was investigated (Wang 2024). The method in turn not only achieved high accuracy but also demonstrated good generalization ability, imparting an efficient support tool for marketing practitioners.

Yet a different method called as multilingual BERT-based classifiers and zero-shot classification was designed to enhance the accuracy and applicability in classification of multi lingual data (Manias *et al*., 2023). A method to ensure robustness using neural models was proposed with improved precision (Tüselmann *et al*., 2024). A

comprehensive review and challenges in deep learning for document classification was designed (Md. Islam *et al*., 2024). A complete review on ML and DL techniques for semantic document classification via speech mechanism was investigated (Tyagi *et al*., 2024).

Yet another deep learning technique to focus on accuracy was presented (Khan *et al*., 2024). A hybrid deep learning mechanism was applied to encapsulate both local and contextual sentiment information in an efficient fashion (Jain *et al*., 2025). However focus was not made in multi-label classification. To focus on this aspect, a fusion of ML and deep learning methods were investigated for both binary and multi-class classification of documents (Uddin *et al*., 2024).

In order to examine the many amounts of texts, researchers are improved confronting the concern of text classification. While the manual labeling is unfeasible researchers have to ascertain automatized mechanisms for text classification. Nevertheless, the show remainder understudied in social sciences. A review of deep learning techniques was presented for accurate classification (Hersh *et al*., 2023).

A holistic comparison of deep learning techniques for short text classification was proposed with much less training time and resources (Shyrokykh *et al*., 2023). Machine learning algorithms were applied for multi-class classification of documents containing COVID 19 cases (Rabby *et al*., 2022). The main objective here remained in ascertain the information type and representation method that influence biomedical document classification task both accurately and precisely. The evaluation of DL techniques for article level classification was investigated (Rivest *et al*., 2021). A comprehensive survey of document classification techniques employing artificial intelligence was presented to highlight experimental insights (Taha *et al*., 2024). Yet another review of deep learning techniques for complex document classification focusing on data annotation and class imbalance was designed (Jamieson *et al*., 2024).

Contributions of the work

Major contribution of this manuscript is that we introduce Lancaster Tokenized Sliced Inverse Regressive Keyword Extraction (LT-SIRKE) for social media documents to investigate semantic analysis user satisfaction level towards improved understanding of data. The key contributions include the following:

- To propose a method called Lancaster Tokenized Sliced Inverse Regressive Keyword Extraction (LT-SIRKE) for social media documents with the intent of investigating semantic analysis towards enhanced search and information retrieval significantly.
- To minimize the keyword extraction time and reduce the error involved in analyzing semantic data, NLP-based Keras Lancaster and Zipf's Law-based Query Pre-processing is designed. Using the Keras-based tokenization, Lancaster-based stemming and Zipf's Law-basis of stop word elimination, model can ensure minimum error and extraction time extensively.
- To present the Bayesian Averaging and Sliced Inverse Regression-based Keyword extraction with the intent of facilitating efficient information access via Gaussian Kernel Bayesian Averaging for improving precision and recall rate involved while investigating semantic analytic results on social media documents extensively.
- Result of proposed technique is also estimated. Simulation results show that the LT-SIRKE method investigate the semantic analytic results on social media documents with minimum error, keyword extraction time and maximum precision, accuracy.

Related works

Natural language processing (NLP) makes certain machines to ascertain and process human language. As far as the NLP field is considered, text analysis is considered as a major area of research where semantic information is derived from textual input data. Over the past few years, analysis of text has made a great deal of attention and is successfully deployed in a wide range of real-world applications.

Inferences in sentiment and semantic analysis using machine learning algorithms were investigated in [16]. Yet another handwritten text recognition based on cross-modal segmentation was designed in [17] for robust semantic word representation. A comprehensive review of text analytics employing deep learning techniques was investigated in [18] with the main emphasize on popular publications. Also a conclusion was made that by applying long short term memory there resulted in overall text analytics task performance.

Every day, an extensive amount of structured and unstructured data is created, however, it remains unanalyzed. In prevailing industrial background, proportional amount of sectors are struggling through issues brought up in unstructured information that in turn causes huge financial losses totaling to millions annually. If put to use efficiently, this data has the prospective to considerably improve operational efficiency in an extensive manner.

A comprehensive review of analysis of semantic speech employing machine learning technique was investigated in [19]. Yet another systematic review on NLP on biological data employing machine learning was presented in [20]. Unstructured document analysis employing AI-driven methods to center both on accuracy and precision facets was presented [21]. Over the past few years document analysis has received its popularity owing to its easy storage and retrieval. However, accessing them is found to be both laborious and cumbersome process. To address on this issue, exploring sentiment analysis for hand written documents employing advanced machine learning algorithms was

presented in [22]. By employing advanced machine learning for hand written documents resulted in higher accuracy.

Employing NLP text data were analyzed using dictionary-based approaches [23]. By employing this approach ensured tradeoff between accuracy and interpretability. However with the high level of inclusion of noise error involved was not concentrated. To focus on this aspect, conceptual model in NLP [24] was designed that in turn reduced the error rate significantly. Yet another historical methodological review focusing on the trajectory of semantic analysis was presented in [25]. A transformer based method was proposed in [26] for handling automatic verification of semantic financial documents. By employing this method resulted in the overall enhancement of generating accurate and informative abstract summaries.

The mushroom growth of social media transfigured how people sight connections. ML basis of sentiment analysis as well as classification of news aid in comprehending both emotion and accessing news extensively. Nevertheless, most studies concentrate on complicated models necessitating heavy resources, making deployment both laborious and cumbersome resource-limited environments.

A stochastic gradient descent method employing ridge classifier was proposed in [27] to boost the overall performance. Also, by using ensemble classifier resulted in the overall accuracy improvement. A holistic systematic mapping review on the utilization of NLP-based text representation techniques was investigated in [28]. An exhausting systematic review on the application of machine learning and NLP over a period of twenty years, its merits and demerits were discussed in [29]. A comprehensive survey on machine learning techniques for document analysis was investigated in [30].

Inspired by the above area, in this work a method for investigating semantic analytics on social media data called Lancaster Tokenized Sliced Inverse Regressive Keyword Extraction (LT-SIRKE) is proposed to ensure minimum keyword extraction time, error rate with improved accuracy and precision is presented.

Methodology

Semantic analytics dispenses meaningful opinion extraction by determining emerging concept sets in place of inspecting the occurrences of secluded words. Nevertheless, comprehending and utilizing social media documents in a significant way is still a big issue. In this section a method called, Lancaster Tokenized Sliced Inverse Regressive Keyword Extraction (LT-SIRKE) for conducting significant semantic analysis on social media documents is presented. Figure 1 shows the structure of LT-SIRKE method.

As shown in the above figure, the LT-SIRKE technique is divide as pre-processing ,feature extraction. To start with, the user inputs their query into the user window for efficient processing. The query pre-processing is performed

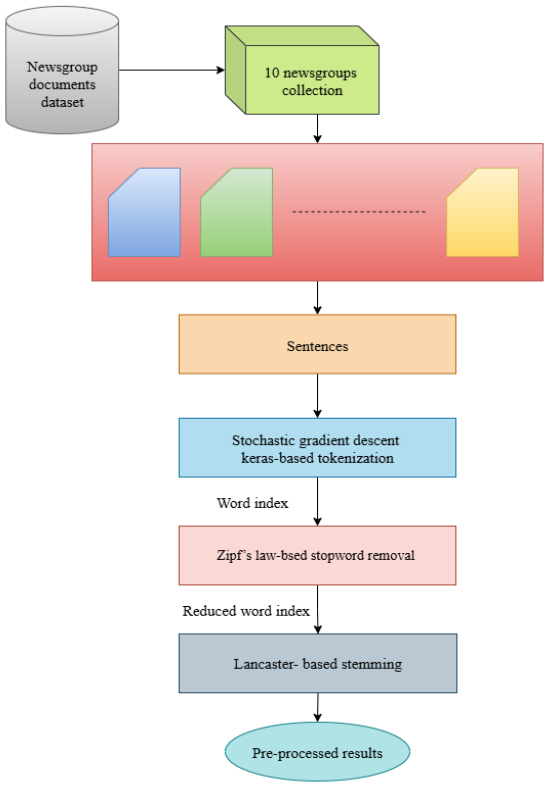

Figure 1: Structure of LT-SIRKE method [updated]

employing Keras Lancaster and Zipf's Law model. Next, with the pre-preprocessed results are provided as input to the Sliced Inverse Regression-based feature extraction model for efficient semantic keyword extraction.

Data collection

The Dataset Text Document Classification employed in the LT-SIRKE method for semantic analytics on social media document is a collection newsgroup documents extracted fromhttps://www.kaggle.com/datasets/jensenbaxter/10dataset-text-document-classification. The database comprises of a file that includes reference to document ID number as well as newsgroup it connected by. Moreover, it includes 10 files that consist of all of the documents with each document per newsgroup. Each novel message at bundled file begins through the four headers. Table 1 given below provides a list of 10 newsgroups in the Dataset Text Document Classification.

Employing the 10 newsgroups in the Dataset Text Document Classification a robust Semantic Analytics on (NLP) Social Media Documents using LT-SIRKE method is designed in the following sub-sections.

NLP-based Keras Lancaster and Zipf's Law-based Query Pre-processing

Social media has been recognized as a paramount identified as an essential open channel of sourcing information. People are found to be inclined to communicate information

Table 1: Newsgroup in Dataset Text Document Classification

S. No	*Newsgroups*
1	Business
2	Entertainment
3	Food
4	Graphics
5	Historical
6	Medical
7	Politics
8	Space
9	Sport
10	Technology

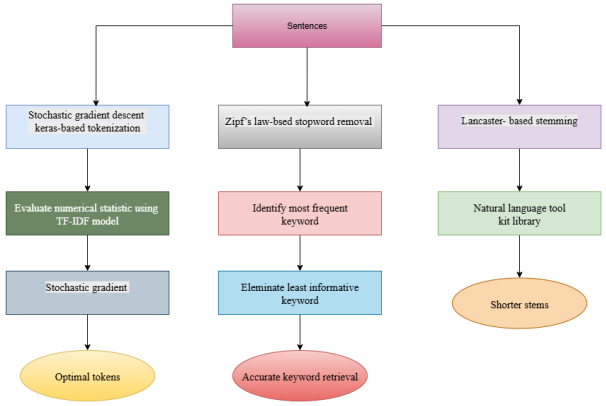

Figure 2: Structure of NLP-based Keras Lancaster and Zipf's Law-based Query Pre-processing model

more frequently and swiftly via social media owing to its real-time nature. Contrary to meticulously generated news and distinct literary contents of the web, social media posts present several new issues for analytics algorithms owing to their noisyand social nature. Majority of the social media contents consists of information term usage, fragmented statements noisy data statement with spelling and grammatical mistakes abbreviated phrases/words and so on. These issues makeit both laborious and cumbersome to seek improvement of the performance of prevailing pre-processing solutions for social media documents. In this section NLP-based Keras Lancaster and Zipf's Law-based Query Pre-processing in NLP is designed. Figure 2 shows the structure of NLP-based Keras Lancaster and Zipf's Law-based Query Pre-processing model.

As shown in the above figure, the NLP-based Keras Lancaster and Zipf's Law-based Query Pre-processing model is split into three parts. They are Stochastic Gradient Descent Keras-based tokenization, Lancaster-based stemming and Zipf's Law-based stop word removal.

Let us consider 10 newsgroups in the Dataset Text Document Classification 'DS' for performing pre-processing. With the dataset 'DS' split into 10 newsgroups the mathematical formulates for generating input matrix is represented as given below.

$$IM = \{NG_1, NG_2, \ldots., NG_N\}, where\ N = 10 \quad (1)$$

$$NG1 = (Business) \rightarrow \{B_1, B_2, \ldots, B_{100}\}; NG2 = (Entertainment) \rightarrow \{E_1, E_2, \ldots, E_{100}\}; NG3 = (Food) \rightarrow \{F_1, F_2, \ldots, F_{100}\}; NG4 = (Graphics) \rightarrow \{G_1, G_2, \ldots, G_{100}\}; NG5 = (Historical) \rightarrow \{H_1, H_2, \ldots, H_{100}\}; NG6 = (Medical) \rightarrow \{M_1, M_2, \ldots, M_{100}\}; NG7 = (Politics) \rightarrow \{P_1, P_2, \ldots, P_{100}\}; NG8 = (Space) \rightarrow \{S_1, S_2, \ldots, S_{100}\}; NG9 = (Sport) \rightarrow \{SP_1, SP_2, \ldots, SP_{100}\};\ NG10 = (Technology) \rightarrow \{T_1, T_2, \ldots, T_{100}\} \quad (2)$$

From the above equations (1) and (2), 'NG_N' represents 'N' newsgroups with each newsgroups split according to 10 different newsgroups, i.e., business 'B', entertainment 'E', food 'F', graphics 'G', historical 'H', medical 'M', politics 'P', space 'S', sport 'SP' and technology 'T'. Also each newsgroup'scomprises of 100 files.

Initially in LT-SIRKE method, the user inputs their query into the user window. Afterward, the query is sent to the system for efficient processing. Then, query pre-processing is carried out to optimize search effectiveness as well as minimize computational time. In query preprocessing phase, tokenization, stemming as well as stop word elimination process is performed. Stochastic Gradient Descent Keras tokenizer is utilized in LT-SIRKE method to segment the input query into individual words or tokens.Stochastic Gradient Descent (SGD) Keras tokenizer optimizes models by minimizing errors. The SGD being a variation of gradient descent utilizes small batches of data to fine-tune model parameters, therefore reducing the error involved in semantic analytics on social media data.

To start with the dataset is initially tokenizes with retaining only 5000 files and then converting training and testing to the sequence of matrices. Also the training and testing labels are converted categorically to having a total of 10 newsgroups. Now the TF-IDF model of tokenizer is employed to perform tokenization and is mathematically formulated as given below.

$$tf\left(t, NG\right) = \frac{freq_{t,NG}}{\sum_{t' \in NG} freq_{t',NG}}\left[IM\right] \quad (3)$$

$$idf\left(t, NG\right) = \log \frac{M}{\left|\left\{NG \in DS; t \in NG\right\}\right|}\left[IM\right] \quad (4)$$

From the above equation (3), the term (i.e. text) frequency 'tf' with respect to text '

t' and newsgroup '
NG' is formulated based on the frequency '$freq_{t,NG}$' of text '
t' in newsgroup '
NG' and the number of newsgroup '
NG' containing text '
t', i.e. '
$freq_{t',NG}$' respectively. In addition from the above equation (4), the inverse social media document frequency 'idf' results are arrived at based on the total number of documents (i.e. files) in the newsgroup 'M' respectively. Now with the stochastic gradient let us minimize an objective function (i.e. an error function) as given below.

$$Q\left(tf\left(t,NG\right)\right)=\frac{1}{n}\sum_{i=1}^{n}Q_i\left(tf\left(t,NG\right)\right) \tag{5}$$

From the above equation (5) the parameter '
$tf\left(t,NG\right)$' that minimizes '$Q\left(tf\left(t,NG\right)\right)$' is evaluated. By employing this stochastic gradient function for tokenizing the error is said to be reduced extensively. The procedure of eliminating affixes from a word with the objective that it is left with the stem of that word is referred to as stemming. For example, consider the words 'walk', 'walking' and 'walks'. Upon accomplishment of stemming all are converted into the root word 'walk'. In our work Lancaster-based stemming is used that significantly reduces the number of words, therefore not only lowering the complexity of the sample space but also improves the accuracy as it has not to deal with inflected word forms. The words in tokenized samples are then stemmed using Lancaster stemmer as given below.

$$IM\left[S\right]\rightarrow IM\left[Played;Playings;Plays\right]\rightarrow Play \tag{6}$$

From the above formulate (6) the words in tokenized samples '$IM[S]$' are applied with Lancaster stemmer and accordingly the stemmed results are obtained for further processing. Finally, in the pre-processing step, Zipf's Law-based stop word removal is performed to eliminate high-frequency words carrying little semantic meaning for enhanced retrieval efficiency on social media documents. According to this Zipf's Law the word frequency in a social media document is said to be inversely proportional to its frequency table rank.

To be more specific, small number of words occurs very infrequently while a large number of words occur frequently. Hence, large numbers of words that occur frequently are discarded from the social media documents whereas the small numbers of words that occur infrequently are retained for further processing, therefore improving overall accuracy and efficiency in an extensive manner. These large numbers of words that occur frequently are treated as stop words and removed from index whereasthe small numbers of words that occur infrequently are retained from further processing. Also the advantage of employing Zipf's Law-based stop word removal is that the decision regarding removal or retaining for further processing is made at matching time.

During matching process the weight is decreased upon successful matching and increased on contrary. The Zipf's Law-based stop word removal function is then mathematically formulated as given below.

$$Res=\sum_{t}TF_{t,Q}.\frac{TF_{t,NG}}{TF_{t,NG}+\frac{\mu\left(NG\right)}{\sigma\left(NG\right)}}*\log\left(\frac{\left|C\right|}{DF_t}\right) \tag{7}$$

From the above equation (7) the Zipf's Law-based stop word removal function '$TF_{t,Q}$' is formulated based on the term frequency (i.e. specific text within a social media document) that makes decisions at query (matching time) 'Q' with respect to a text 't' for the corresponding newsgroup '
NG'. Also '$\mu\left(NG\right)$' and '
$\sigma\left(NG\right)$' of the text in the news group is measured along with the 'DF_t' denoting the number of occurrences of a specific text within a newsgroup what is then ranked by frequency and a constant 'C' correlating the frequency of text. The pseudo code representation of NLP-based Keras Lancaster and Zipf's Law-based Query Pre-processing is given below.

Input: Dataset '
DS', Newsgroups '$NG=\left\{NG_1,NG_2,\ldots,NG_N\right\}$'
Output: Computationally-efficient Pre-processed Text '
PT' from different newsgroups

1: Initialize '$N=10$', Weight '$w=0$'
2: Begin
3: For each Dataset '
DS' with Newsgroups 'NG'
4: Formulate input matrix according to (1) and (2)
// Stochastic Gradient Descent Keras-based tokenization
5: Obtain Stochastic Gradient Descent Keras based tokenized results according to (3) and (4)
6: Formulate stochastic gradient objective function according to (5)
7: Return error-minimized tokenized results
//Lancaster-based stemming
8: Formulate stemmed results according to (6)
//Zipf's Law-based stop word removal
9: Evaluate Zipf's Law-based stop word removal function according to (7)
10: If '$Res\geq 0$'
11: Then occurrence of small numbers of text
12: Then '$w=w+1$'
13: Text are retained for further processing '
PT'
14: Return text from corresponding newsgroup
15: End if

16: If'
$Res < 0$'
17: Thenoccurrence of large numbers of text
18: Then'
$w = w - 1$'
19: Text are discarded
20: Go to step 4
21: End if
22: End for
23: End

Algorithm 1 NLP-based Keras Lancaster and Zipf's Law-based Query Pre-processing

As given in the above algorithm, the NLP-based Keras Lancaster and Zipf's Law-based Query Pre-processing is split into three parts. First, with the query obtained from the user into the user window is subjected to Keras-based tokenization or Stochastic Gradient Descent (SGD) Keras tokenizer. By applying this tokenizer along with the SGD function optimizes tokenizing process by minimizing errors extensively, therefore reducing the overall error rate involved in Keyword Extraction-based Semantic Analytics on Social Media Documents. Second Lancaster-based stemming is applied to the tokenized samples that using the stemming function reduces words to their root forms, therefore reducing the keyword extraction time. Finally, Law-based stop word removal is applied to the tokenized and stemmed results that by employing the Zipf's Law perform stop word removal process that by making the decision only during the query or matching time aids in the improvement of overall accuracy in an extensive manner.

Bayesian Averaging and Sliced Inverse Regression-based Keyword Extraction

Keyword Extraction (KE) is a paramount and an important and knowledgeable problem as far as NLP is concerned with applications varying from recommending academic papers, online advertising, clustering websites and so on. Though Keyword Extraction has predominantly been executed in the area of academic papers, in this work it is carried out in the domain of social media. Social media documents an important genre for managing social media activities. Several of us are frequently faced with the challenging task of increasingly large amounts of social listening insights on a daily basis. Keywords extracted from social media documents can assist us fight against such information overload by permitting a structured investigation of the newsgroups contained in social media documents. Existing literature on keyword extraction has not balanced the social media document email genre.Bayesian Averaging as well as Sliced Inverse Regression-based Keyword Extraction to facilitate information access is proposed. Figure 3 shows the structure of Bayesian Averaging and Sliced Inverse

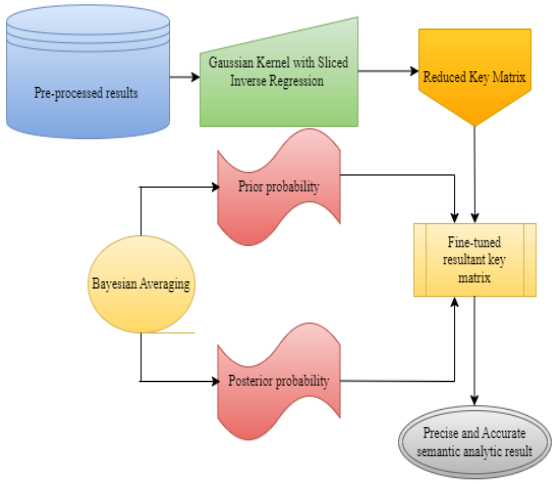

Figure 3: Structure of Bayesian Averaging and Sliced Inverse Regression-based Keyword Extraction model

Regression-based Keyword Extraction for accurate and precise information access.

As shown in the above figure, with the pre-processed samples as input, initially, Gaussian Kernel function with Sliced Inverse Regression is applied to obtain reduced keyword matrix. Following which Bayesian Averaging function is applied separately to obtain both prior and posterior probability results, therefore obtaining fine-tuned resultant key matrix both precisely and accurately. Let us select the Gaussian Kernel function with the mathematical formulate as given below.

$$W\left(NG_i, NG_j\right) = \exp\left(-\frac{\left|NG_j - NG_i\right|^2}{2\sigma_b^2}\right) \quad (8)$$

$$\left|NG_j - NG_i\right| = \sqrt{\left(NG_{j1} - NG_{i1}\right)^2 + \left(NG_{j2} - NG_{i2}\right)^2 + \ldots + \left(NG_{jn} - NG_{in}\right)^2} \quad (9)$$

From the above equations (8) and (9), 'NG_i' and 'NG_j' denotes the input vectors representing newsgroup factor vectors selected from 'NG'. On the other hand '$\left|NG_j - NG_i\right|^2$' denotes the square Euclidean distance between 'NG_i' and 'NG_j' respectively. Moreover, 'σ_b' represents the Gaussian kernel bandwidth factor with larger the 'σ_b' the wider the kernel reach and on contrary smaller the 'σ_b' results in more localized function. Also '$W\left(NG_i, NG_j\right)$' signifies the similarity between newsgroups 'NG_i' and 'NG_j' with the value of '$W\left(NG_i, NG_j\right)$' increasing as they grow closer and the value of '$W\left(NG_i, NG_j\right)$' approaching zero when they are distant.

Employing Gaussian Kernel averaging, the average function within the newsgroup for 'NG_j' is measured where the local mean is evaluating by weighting the 'k' nearest neighbors with the Gaussian kernel function as given below.

$$Z = \mu_i, local = \frac{\sum_{j \in k} NG_j W\left(NG_i, NG_j\right)}{\sum_{j \in k} W\left(NG_i, NG_j\right)} \quad (10)$$

From the above equation (10), 'k' denotes the nearest neighbor keyword index set for data point or newsgroup 'NG_i' with the weights '$W\left(NG_i, NG_j\right)$' measured using the Gaussian kernel function. Using 'β' the original feature or the newsgroup matrix 'NG' is transformed into reduced feature or reduced keyword matrix 'RK' as given below.

$$RK = \beta^T NG, where\, \beta = \left[\omega_1, \omega_2, \ldots, \omega_k\right] \quad (11)$$

From the above equation (11) having obtained the dimensionally reduced keyword matrix 'RK', the social media document dataset now efficiently isolates crucial information for semantic analytics via sliced inverse regression, notably minimizing the original newsgroup dimensionality. This not only improves overall data processing efficiency but also enhances the methods accuracy in facilitating efficient information access. On the basis ofthis results Bayesian Averaging is constructed utilizing the dimensionally reduced keyword matrix 'RK' and the corresponding response variable 'Y'. Let us define '$Prob(Y \mid RK, m)$' are as the likelihood function of 'Y' given model 'm' and the dimensionally reduced keyword matrix 'RK'. Then, the likelihood function '$Prob(Y \mid RK, \gamma_m, m)$' is defined as the probability of observing response variable 'Y' given 'RK', 'γ_m' and 'm' as given below.

$$Prob\left(Y \mid RK, m\right) = \int Prob\left(Y \mid RK, \gamma_m, m\right) Prob\left(\gamma_m \mid m\right) d\gamma_m \quad (12)$$

From the ave equation (12), '$Prob\left(\gamma_m \mid m\right)$' denotes the prior distribution of 'γ_m'. Also it is obvious that the marginal likelihood function '$Prob\left(Y \mid RK, m\right)$' not only depend on specific parameter values 'γ_m' but also is to obtain its posterior probability '$Prob\left(m \mid Z, RK\right)$' as given below.

$$P = Prob\left(Y \mid RK, m\right) Prob\left(m\right) \quad (13)$$

From the above equation (13) '$Prob\left(m\right)$' denote the prior probability and the posterior probability '$Prob\left(m \mid Z, RK\right)$' representing weights in the model averaging respectively. Then, for a new observation (i.e., the test set) 'RK_{update}', weighted average prediction is measured utilizing the Bayesian Averaging. The Bayesian Averaging is the weighted average of all predictions, with weights given by each model's posterior probability '$Prob\left(m \mid Z, RK\right)$' yielding the prediction results '$Y'\left(RK_{new}\right)$'

$$Y'\left(RK_{new}\right) = \sum_m Prob\left(m \mid Z, RK\right) Y_m^{'}\left(RK_{new}\right) \quad (14)$$

From the above equation (14), '$Y'\left(RK_{new}\right)$' denotes the semantic analytics prediction results of model 'm' for the test set 'RK_{new}' respectively. The pseudo code representation of Bayesian Averaging and Sliced Inverse Regression-based Keyword Extraction is given below.

Input: Dataset 'DS', Newsgroups '$NG = \left\{NG_1, NG_2, \ldots, NG_N\right\}$'

Output: Precise semantic analytic results on social media documents

1: **Initialize**Pre-processed Text 'PT' from different newsgroups
2: **Begin**
3: **For** each Dataset 'DS' with Newsgroups 'NG' and pre-processed Text 'PT.' from different newsgroups
4: Formulate Gaussian Kernel function for each Newsgroups 'NG.' according to (8) and (9)
5: Evaluate Gaussian Kernel averaging according to (10)
6: Obtain reduced keyword matrix 'RK.' according to (11)
7: Obtain likelihood function according to (12)
8: Evaluate marginal likelihood function according to (13)
9: Obtain the semantic analytic prediction results according to (14)
10: **Return** fine-tuned reduced keyword results '$Y'\left(RK_{new}\right)$'
11: **End for**
12: **En**d

Algorithm 2 Bayesian Averaging and Sliced Inverse Regression-based Keyword Extraction

As given in the above algorithm, two main processes are performed during the keyword extraction process, namely, the application of Sliced Inverse Regression and the fine-tuning of keyword results employing Bayesian Averaging. Employing Sliced Inverse Regression function to social media documents assist in identifying the most relevant linear combinations of keywords while reducing the complexity of data, therefore improving the overall prediction rate. Also by exploiting Bayesian Averaging function on Social Media documents involving combining prediction from multiple newsgroups generate a more robust and accurate outcome extensively.

Experimental Setup

The performance of the Lancaster Tokenized Sliced Inverse Regressive Keyword Extraction (LT-SIRKE) methods is estimated through comparing forecast by actual data with Python language. Evaluation metrics used included: (1) prediction accuracy or simply: The percentage of correct semantic analytic predictions on social media data out of the total predictions, (2) precision: the proportion of true positive semantic analytic predictions on social media data out of the total positive predictions, (3) error or prediction error: the proportion of erroneous semantic analytic prediction results out of the total samples, (4) keyword extraction time: the time consumed in investigating the extraction time

involved in the keyword generation process and (5) recall: the proportion of true positive semantic analytic predictions on social media data out of total negative predictions.

Through evaluation stage, five performance parameters were computed for every technique to find out their effectiveness for semantic analysis on social media data. Complete outcomes and comparisons are presented in succeeding part on Results and Discussion. Three techniques were tested to communicate complete viewpoint on numerous methods in machine learning. These techniques were chosen to calculate effectiveness of dissimilar ML techniques in semantic analysis on social media data, covering as of existing methods to newest conventional comprising, BERT-MultiLayered Convolutional Neural Network (B-MLCNN) [1] and Latent Dirichl*et al*location (LDA) [2] along with two state-of-the-art methods, BERT and LSTM.

Case study and Inferences

In this section case analysis of semantic analysis on social media document using 10 newsgroups collection Dataset are simulated by applying method called, Lancaster Tokenized Sliced Inverse Regressive Keyword Extraction (LT-SIRKE). With an overall of 10 newsgroups collection present in the dataset, in our work, food newsgroup of 11th document is analyzed. Figure 4 given below shows the input set.

Food_11.txt input text

3 1/4	cups all-purpose flour
2 1/2	teaspoons baking powder
1/2	cup butter or margarine -- softened
1	cup sugar
3	eggs
2	teaspoons lemon peel -- finely shredded
1/4	teaspoon almond extract
1	pinch saffron -- if desired
1/2	cup almonds -- finely chopped,toast
1	egg white

Combine flour and baking powder. In large mixer bowl beat butter and sugar until blended. Beat in eggs, lemon peel, almond extract and saffron. Beat in flour mixture until well blended. Stir in almonds. Divide dough in half. Shape each portion into a 12x2x1-inch loaf. Place 6 inches apart on a lightly greased cookie sheet. Beat the egg white until foamy. Brush over tops of loaves. Bake in 375F oven 20 to 25 minutes or until light brown. Cool on cookie sheet about 1 hour. Cut each loaf diagonally into 1/2- inch thick slices. Lay slices, cut side down, on cookie sheet. Bake in a 325F oven 10 minutes longer or until dry and crisp. Cool on wire rack. These cookies are good made several days ahead and stored in a paper bag to soften slightly. To store longer, place in a covered container. Makes about 36 cookies.

Figure 4: Food 11 Input text

Food_11.txt

3	1/4	cups	all-purpose	flour	2	1/2	teaspoons
baking	powder	1/2	cup	butter	or	margarine	--
softened	1	cup	sugar	3	eggs	2	teaspoons
lemon	peel	--	finely	shredded	1/4	teaspoon	almond
extract	1	pinch	saffron	--	if	desired	1/2
cup	almonds	--	finely	chopped,toast	1	egg	white
Combine	flour	and	baking	powder.	In	large	mixer
bowl	beat	butter	and	sugar	until	blended.	Beat
in	eggs,	lemon	peel,	almond	extract	and	saffron.
Beat	in	flour	mixture	until	well	blended.	Stir
in	almonds.	Divide	dough	in	half.	Shape	each
portion	into	a	12x2x1-inch	loaf.	Place	6	inches
apart	on	a	lightly	greased	cookie	sheet.	Beat
the	egg	white	until	foamy.	Brush	over	tops
of	loaves.	Bake	in	375F	oven	20	to
25	minutes	or	until	light	brown.	Cool	on
cookie	sheet	about	1	hour.	Cut	each	loaf
diagonally	into	1/2-	inch	thick	slices.	Lay	slices,
cut	side	down,	on	cookie	sheet.	Bake	in
a	325F	oven	10	minutes	longer	or	until
dry	and	crisp.	Cool	on	wire	rack.	These
cookies	are	good	made	several	days	ahead	and
stored	in	a	paper	bag	to	soften	slightly.
To	store	longer,	place	in	a	covered	container.
Makes	about	36	cookies.				

Figure 5: Tokenized results

The result of Stochastic Gradient Descent Keras-based tokenization is shown in Figure 5.

Following which Lancaster-based stemming is applied to the tokenized results and results are illustrated in Figure 6.

Next, Zipf's Law-based stop word removal process is carried out with the tokenized and stemmed results to finally generate words after removal of stop words as shown in Figure 7.

Finally by combining these results a pre-processed result or reduced text are generated for further processing. By applying Stochastic Gradient Descent (SGD) being a variation of gradient descent employs small batches of data to fine-tune model parameters that in turn aids in minimizing the error rate. Also using Lancaster-based stemming along with the tokenized samples in turn reduces words to their

Food_11.txt Lancaster-based stemmed words					
Word	Lancaster Stem	Word	Lancaster Stem	Word	Lancaster Stem
cups	cup	combine	combin	oven	ov
all	al	large	larg	minutes	minut
purpose	purpos	mixer	mix	diagonally	diagon
flour	flo	blended	blend	slices	slic
teaspoons	teaspoon	mixture	mixt	side	sid
baking	bak	well	wel	longer	long
powder	powd	divide	divid	wire	wir
butter	but	shape	shap	these	thes
margarine	margarin	portion	port	cookies	cooky
softened	soft	place	plac	are	ar
sugar	sug	inches	inch	made	mad
eggs	eg	lightly	light	several	sev
finely	fin	greased	greas	days	day
desired	desir	cookie	cooky	stored	stor
almonds	almond	over	ov	paper	pap
chopped	chop	tops	top	soften	soft
egg	eg	loaves	loav	slightly	slight
white	whit	bake	bak	store	stor
covered	cov	container	contain	makes	mak

Figure 6: Stemmed results

root forms, hence reducing the overall keyword extraction time of the LT-SIRKE method. Finally, with the reduced keywords as input is subjected to Bayesian Averaging and Sliced Inverse Regression-based Keyword Extraction model. Here to extract keywords Bayesian Averaging in addition to Sliced Inverse Regression function is applied. The results are shown in figure 8.

By applying Sliced Inverse Regression function and reversing regression approach in turn aids in examining how the features relate to response variable via inverse relationship. This in turn extracts lower dimensional data representation while preserving the most predominant information, therefore improving overall precision and recall. Also by using Bayesian Averaging robust and accurate feature representation is ensured. The elaborate quantitative analysis is provided in the following sub-sections.

Results

Results present experiential findings as well as case studies showcasing effectiveness of Lancaster Tokenized Sliced Inverse Regressive Keyword Extraction (LT-SIRKE) for semantic analytics on social media data with main focus on keyword extraction. During quantitative and qualitative study of customer satisfaction level parameters, this part aspire to give insights to tangible advantages and manipulate of these techniques on entire design procedure. Experiential findings as well as case studies offer concrete verification of how LT-SIRKE with two existing methods, BERT-MultiLayered Convolutional Neural Network (B-MLCNN) [1] and Latent Dirichl*et al*location (LDA) [2] along with two state-of-the-art methods, BERT and LSTM have contributed to investigating semantic analytics on social media data across different domains.

Precision, recall and accuracy

To start with precision employing three distinct methods, LT-SIRKE, B-MLCNN [1] and LDA [2],along with two state-of-the-art methods, BERT and LSTM is presented for semantic analytics on social media documents. The rate of precision involved in investigating u semantic analytics on social media documents is provided below.

Food_11.txt input text	Zipf's law-based stopword removed text
3 1/4 cups all-purpose flour 2 1/2 teaspoons baking powder 1/2 cup butter or margarine -- softened 1 cup sugar 3 eggs 2 teaspoons lemon peel -- finely shredded 1/4 teaspoon almond extract 1 pinch saffron -- if desired 1/2 cup almonds -- finely chopped,toast 1 egg white Combine flour and baking powder. In large mixer bowl beat butter and sugar until blended. Beat in eggs, lemon peel, almond extract and saffron. Beat in flour mixture until well blended. Stir in almonds. Divide dough in half. Shape each portion into a 12x2x1-inch loaf. Place 6 inches apart on a lightly greased cookie sheet. Beat the egg white until foamy. Brush over tops of loaves. Bake in 375F oven 20 to 25 minutes or until light brown. Cool on cookie sheet about 1 hour. Cut each loaf diagonally into 1/2- inch thick slices. Lay slices, cut side down, on cookie sheet. Bake in a 325F oven 10 minutes longer or until dry and crisp. Cool on wire rack. These cookies are good made several days ahead and stored in a paper bag to soften slightly. To store longer, place in a covered container. Makes about 36 cookies.	Removed Words: {'into', 'in', 'to', 'of', 'a', 'and', 'or', 'on'} Cleaned Text: 3 1 4 cups all purpose flour 2 1 2 teaspoons baking powder 1 2 cup butter margarine softened 1 cup sugar 3 eggs 2 teaspoons lemon peel finely shredded 1 4 teaspoon almond extract 1 pinch saffron if desired 1 2 cup almonds finely chopped toast 1 egg white combine flour baking powder large mixer bowl beat butter sugar until blended beat eggs lemon peel almond extract saffron beat flour mixture until well blended stir almonds divide dough half shape each portion 12x2x1 inch loaf place 6 inches apart lightly greased cookie sheet beat the egg white until foamy brush over tops loaves bake 375f oven 20 25 minutes until light brown cool cookie sheet about 1 hour cut each loaf diagonally 1 2 inch thick slices lay slices cut side down cookie sheet bake 325f oven 10 minutes longer until dry crisp cool wire rack these cookies are good made several days ahead stored paper bag soften slightly store longer place covered container makes about 36 cookies

Figure 7: Stop word removed results

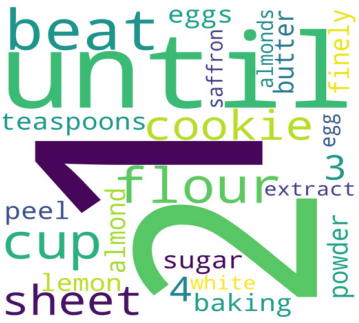

Figure 8: Keywords extracted using Sliced Inverse Regression and Bayesian Averaging

$$Pre = \frac{TP}{TP + FP} \tag{15}$$

$$Rec = \frac{TP}{TP + FN} \tag{16}$$

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{17}$$

From the above equations (15), (16) and (17), precision 'Pre', recall 'Rec' and accuracy 'Acc' is calculated depend on the true positive 'TP' (i.e. keywords extracted from business newsgroup returned as business newsgroup), 'TN' (i.e. keywords extracted from entertainment newsgroup returned as entertain newsgroup), 'FP' (i.e. keywords extracted from business newsgroup returned as entertain newsgroup) and false negative rate 'FN' (i.e. keywords extracted from entertainment newsgroup returned as business newsgroup) respectively. Higher precision, recall and accuracy rate ensures the efficiency of the method in providing semantic analytic results and vice versa. Table 2 presents a comparison of 'Pre', 'Rec' and 'Acc' factors for the proposed, LT-SIRKE, B-MLCNN [1] and LDA [2] with BERT and LSTM depend on dataset text document dataset.

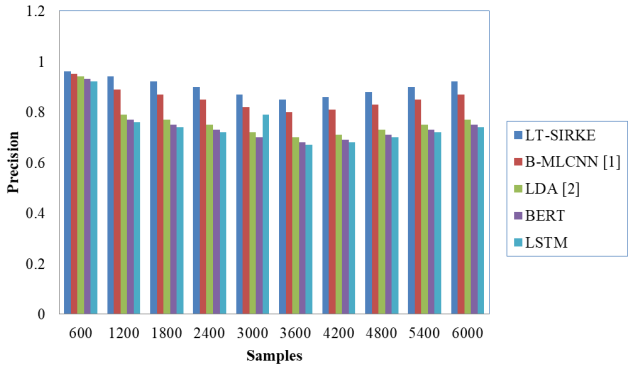

Figure 9: Precision versus samples

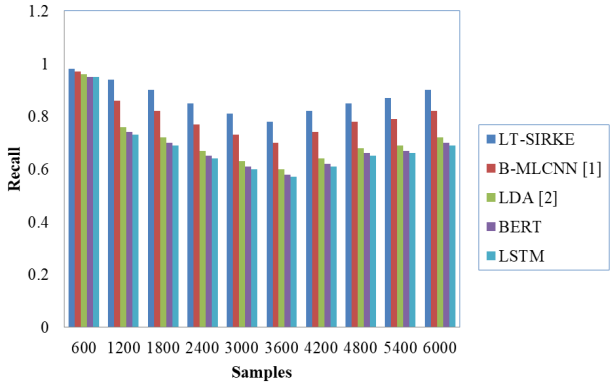

Figure 10: Recall versus samples

Figure 9 given above shows the graphical representation of precision versus 6000 different samples obtained from 10 different newsgroups. From the above figure, the precision rate of LT-SIRKE method was found to be better upon comparable to [1], [2] with two state-of-the-art methods, BERT and LSTM.

Figure 10 given above illustrates graphical representation of recall with respect to 6000 different samples collected

Table 2: Impact of precision, recall and accuracy on Dataset Text Document Classification using LT-SIRKE and two existing methods, B-MLCNN [1], LDA [2], BERT and LSTM[updated]

Samples	*Precision*					*Recall*					*Accuracy*				
	LT-SIRKE	*B-MLCNN [1]*	*LDA [2]*	*BERT*	*LSTM*	*LT-SIRKE*	*B-MLCNN [1]*	*LDA [2]*	*BERT*	*LSTM*	*LT-SIRKE*	*B-MLCNN [1]*	*LDA [2]*	*BERT*	*LSTM*
600	0.96	0.95	0.94	0.93	0.92	0.98	0.97	0.96	0.95	0.95	0.95	0.93	0.91	0.9	0.89
1200	0.94	0.89	0.79	0.77	0.76	0.94	0.86	0.76	0.74	0.73	0.91	0.86	0.76	0.74	0.73
1800	0.92	0.87	0.77	0.75	0.74	0.9	0.82	0.72	0.7	0.69	0.9	0.85	0.75	0.73	0.72
2400	0.9	0.85	0.75	0.73	0.72	0.85	0.77	0.67	0.65	0.64	0.87	0.82	0.72	0.7	0.69
3000	0.87	0.82	0.72	0.7	0.79	0.81	0.73	0.63	0.61	0.6	0.9	0.85	0.75	0.73	0.72
3600	0.85	0.8	0.7	0.68	0.67	0.78	0.7	0.6	0.58	0.57	0.92	0.87	0.77	0.75	0.74
4200	0.86	0.81	0.71	0.69	0.68	0.82	0.74	0.64	0.62	0.61	0.93	0.88	0.78	0.76	0.75
4800	0.88	0.83	0.73	0.71	0.7	0.85	0.78	0.68	0.66	0.65	0.89	0.84	0.74	0.72	0.71
5400	0.9	0.85	0.75	0.73	0.72	0.87	0.79	0.69	0.67	0.66	0.91	0.86	0.76	0.74	0.73
6000	0.92	0.87	0.77	0.75	0.74	0.9	0.82	0.72	0.7	0.69	0.94	0.89	0.79	0.77	0.76

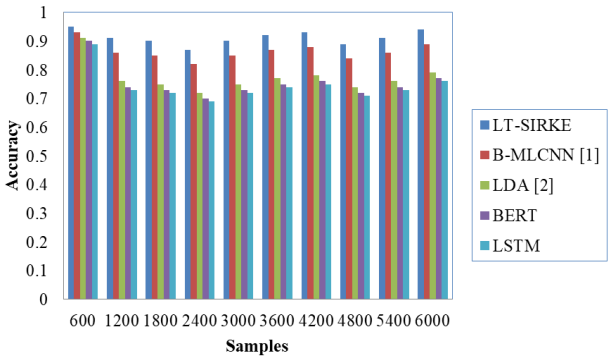

Figure 11: Versus samples

from 10 different newsgroups at different time intervals. From the above figure, recall rate of LT-SIRKE method was found to be better compared to [1], [2], with two state-of-the-art methods, BERT and LSTM.

Figure 11 depicts graphical illustration of involved in semantic analysis on social media datasets with respect to 6000 different. In figure 11, the accuracy rate of LT-SIRKE method was found to be better compared upon comparison to [1], [2], BERT and LSTM. The reasons behind the precision, recall and accuracy improvement using the proposed LT-SIRKE method to be comparatively better than [1], [2], BERT and LSTM could be contributed to the application of Bayesian Averaging and Sliced Inverse Regression-based Keyword Extraction for accurate and precise information access. By applying this keyword extraction algorithm, two main processes were performed while performing keyword extraction process. They are the application of Sliced Inverse Regression and then the fine-tuning of keyword results using Bayesian Averaging function. Using Sliced Inverse Regression function to social media documents aids in identifying the most relevant linear combinations of keywords while minimizing the data complexity considerably. This in turn improves the overall precision rate of the LT-SIRKE by 5% ,15% and 17% than the BERT and LSTM. Also by using Bayesian Averaging function on Social Media documents ensures robust and accurately prediction results from multiple newsgroups, therefore improving the overall recall rate of LT-SIRKE by 8% ,19% than the [1],[2], 21% compared to BERTand 22% compared to LSTM. Finally by applying the Law-based stop word removal in the pre-processing stage that by employing the Zipf's Law analyzes and validates stop word removal procedures that by performing decision matching only during query or matching time aids in the improvement of overall accuracy using the LT-SIRKE by 5% , 15% than the [1],[2], 17% compared to BERTand 18% compared to LSTM.

Error rate

Error rate involved during semantic analytics on social media data. It is formulated as.

Table 3: Impact of error rate on Dataset Text Document Classification using LT-SIRKE and two existing methods, B-MLCNN [1], LDA [2], BERT and LSTM

Samples	*Error rate (%)*				
	LT-SIRKE	*B-MLCNN [1]*	*LDA [2]*	*BERT*	*LSTM*
600	2.83	4.16	5	2.83	4.16
1200	3.15	4.35	5.25	3.15	4.35
1800	3.35	4.85	5.55	3.35	4.85
2400	3.85	5	5.85	3.85	5
3000	4	5.15	6	4	5.15
3600	4.25	5.35	6.15	4.25	5.35
4200	4.55	5.85	6.25	4.55	5.85
4800	4.15	5.25	6	4.15	5.25
5400	3.95	5.05	5.75	3.95	5.05
6000	3.75	4.85	5.35	3.75	4.85

$$ER = \sum_{i=1}^{M} \frac{IM_{IAA}}{IM_i} \quad (18)$$

In equation (18), error rate '' is calculated depend on input matrix containing the samples from 10 different newsgroups '' and the input matrix inaccurately analyzed ''. Minimum the ER more effective the technique is said to be and vice versa. ER is calculated in percentage (%).Table 3 presents a comparison of error rate factor for the proposed, LT-SIRKE, B-MLCNN [1], LDA [2], with BERT and LSTM based on dataset text document dataset.

Figure 12 depicts graphical representation of ER in the vertical axis using the three methods, LT-SIRKE and two existing methods, B-MLCNN [1], LDA [2], BERT and LSTM with respect to 6000 distinct samples provided as input in the horizontal axis. To ensure fair comparison same dataset was employed for all the three methods and the samples were provided as input. From the above graphical representation the error rate were found to be neither increasingly nor decreasingly proportionate to the samples provided as input. This corroborates the aim which enhancing samples not profound results or impacts on ER. However experimental simulation using the three methods found comparatively better results when applied

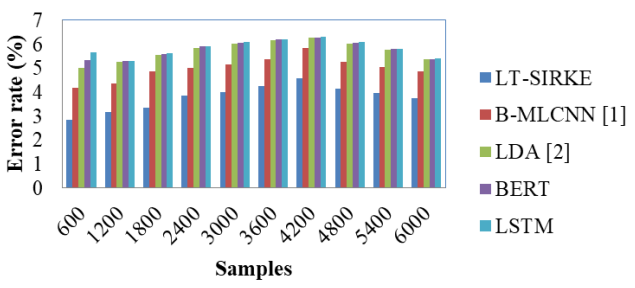

Figure 12: ER versus samples

with LT-SIRKE upon comparison to [1], [2], BERT and LSTM. Cause behind reduction of error rate with the LT-SIRKE was owing to relevance of Stochastic Gradient Descent (SGD) Keras tokenizer in the pre-processing stage. By applying this Stochastic Gradient Descent Keras tokenizer segments the input query into individual words or tokens. Stochastic Gradient Descent (SGD) being a variation of gradient descent employs small batches of data to fine-tune model parameters, therefore minimizing the error rate involved in semantic analytics on social media data using the LT-SIRKE by 33% ,53% than the [1], [2], 55% compared to BERT and 56%compared to LSTM.

Keyword extraction time

Keyword extraction time is measured. To be more specific the keyword extraction time refers to the time consumed in extracting fine tuning keyword matrix. The keyword extraction time is mathematically represented as given below.

$$ET = \sum_{j=1}^{M} IM_j * Tme\left(Y'(RK_{new})\right) \quad (19)$$

In equation (19) extraction time '' is calculated depend on input matrix obtained from '' newsgroup and the actual time consumed in returning fine-tuned resultant keyword '' results. It is measured in terms of seconds (sec). Lower the extraction time more effective the technique is said to be and vice versa. Table 4 presents comparative analysis of ER for the proposed, LT-SIRKE, B-MLCNN [1], LDA [2], BERT and LSTM based on dataset text document dataset.

Figure 13 given above shows the graphical representation of keyword extraction time in the vertical axis and the samples ranging between 600 and 6000 obtained from 10 different newsgroups are provided in the horizontal axis. From the above figure it is incidental enhanicng samples causes an increase in the . But experiments performed at different time intervals for an overall of 10 simulation results

Table 4: Impact of error rate on Dataset Text Document Classification using LT-SIRKE and two existing methods, B-MLCNN [1], LDA [2], BERT and LSTM

Samples	*Extraction time (sec)*				
	LT-SIRKE	*B-MLCNN [1]*	*LDA [2]*	*BERT*	*LSTM*
600	210	246	330	348	378
1200	245	265	315	330	340
1800	275	290	335	350	360
2400	315	335	355	380	390
3000	335	385	405	420	430
3600	350	410	425	450	460
4200	375	435	440	455	465
4800	400	455	465	480	490
5400	415	485	495	510	520
6000	435	510	535	550	560

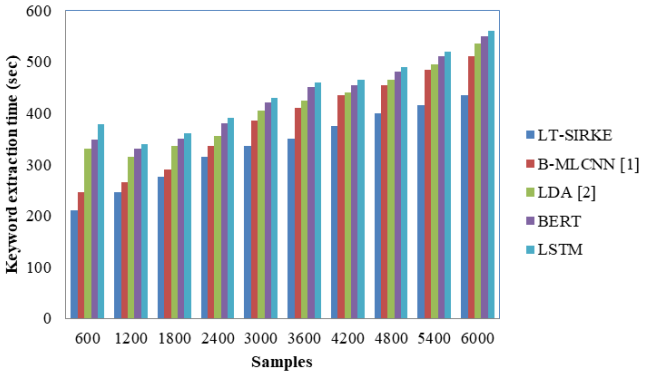

Figure 13: Keyword extraction time versus samples [updated]

shows betterment using proposed LT-SIRKE method upon comparison to [1], [2], BERT and LSTM. The reduction in the keyword extraction time using the proposed LT-SIRKE upon comparison to [1], [2], BERT and LSTMcould be contributed to the application of Lancaster-based stemming. By applying this stemming model, with the tokenized samples being applied with stemming function minimizes words to their root forms, therefore minimizing the keyword extraction time usingproposed LT-SIRKE method by 13% upon comparison to [l], 24% upon comparison to [2], 29% compared to BERTand 33% compared to LSTM.

Conclusion

Semantic analytics employing NLP on social media documents assists computers comprehend the context and meaning of text, resulting in better insights and decision. Past research works underscore semantic analytics with dissimilar conventional as well as non-conventional techniques, using ML, DL. Lancaster Tokenized Sliced Inverse Regressive Keyword Extraction (LT-SIRKE) for social media documents is introduced. Individual process connected through investigation of semantic analytics on social media documents are pre-processing and feature extraction. First, the samples acquired from Dataset Text Document Classification are provided as input to the pre-processing model to perform Keras-based tokenization, Lancaster-based stemming and Zipf's Law-based stop word removal for further processing. Second, Bayesian Averaging and Sliced Inverse Regression-based Keyword Extraction for accurate and precise information access are presented. The proposed LT-SIRKE was experimented in Python language using database Text Document Classification. Experimentation outcomes confirmed LT-SIRKE imparts enhanced outcomes in precision, recall, accuracy, keyword extraction time and error rate compared to the conventional methods and two state-of-the-art methods.

References

Alghamdia, J., Lina, Y. & Luo, S. (2024). Unveiling the hidden patterns: A novel semantic deep learning approach to fake

news detection on social media. *Engineering Applications of Artificial Intelligence, Elsevier*, 137,1-14. https://doi.org/10.1016/j.engappai.2024.109240

Alshuwaier, F., Areshey, A. & Poon, J. (2022). Applications and Enhancement of Document-Based Sentiment Analysis in Deep learning Methods: Systematic Literature Review. *Intelligent Systems with Applications, Elsevier,* 15, 1-25. https://doi.org/10.1016/j.iswa.2022.200090

Atandoh, P., Zhang, F., Adu-Gyamfi, D., Atandoh, P. H. & Nuhoho, R.E. (2023). Integrated deep learning paradigm for document-based sentiment analysis. *Journal of King Saud University - Computer and Information Sciences, Elsevier*, 35 (7), 1-15. https://doi.org/10.1016/j.jksuci.2023.101578

Baxter. J. (2020): (10) Dataset Text Document Classification https://www.kaggle.com/datasets/jensenbaxter/10dataset-text-document-classification

Choi, V. D. J., Verma, S., Kavita, Chatterjee, P., Ijaz, M. F. & Shafi, J. (2022). A Complete Process of Text Classification System Using State-of-the-Art NLP Model. *Computational Intelligence and Neuroscience, Hindawi,* 2022 (49), 1-26. https://doi.org/10.1155/2022/1883698

Galamiton, N., Bacus, S., Fuentes, N., Ugang, J., Villarosa, R., Wenceslao, C. & Ocampo, L. (2024). Predictive Modeling for Sensitive Social Media Contents Using EntropyFlow Sort and Artificial Neural Networks Initialized by Large Language Models. *International Journal of Computational Intelligence Systems, Springer*, 17 (262), 1-18. https://doi.org/10.1007/s44196-024-00668-5

Hersh, Y. Z. P., Xing, F., Ghosh, D., Hobbs, B. D., Craig, Banaei-Kashani, F., Bowlerl, R. P. & Kechris, K. (2023). Deep learning on graphs for multi-omics classification of COPD. *PLOSONE*, 18 (4), 1-23. https://doi.org/10.1371/journal.pone.0284563

Hershowitz, B., Hodkiewicz, M., Bikaun, T., Stewart, M. & Liu, W. (2024). Causal knowledge extraction from long text maintenance documents. *Computers in Industry, Elsevier* 161, 1-15. https://doi.org/10.1016/j.compind.2024.104110

Jain, V., Malviya, L. & Anjana .S. (2025). Optimized hybrid deep learning for cross linguistic sentiment analysis: a novel approach. *Journal of Cloud Computing, Springer*, 14 (30), 1-21. https://doi.org/10.1186/s13677-025-00753-w

Jamieson, L., MorenoGarcía, C. F. & Elyan, E. (2024). A review of deep learning methods for digitisation of complex documents and engineering diagrams. *Artificial Intelligence Review, Springer*, 57 (136), 1-37. DOI:10.1007/s10462-024-10779-2

Jung, S. & Ka, S. (2022). GAE-Based Document Embedding Method for Clustering. *IEEE Access*, 10, 130089 – 130096. DOI: 10.1109/ACCESS.2022.3228548

Khan, S., Abbas, H. & Binsawad, M. (2024). Secure semantic search using deep learning in a blockchain-assisted multi-user setting. *Journal of Cloud Computing: Advances, Systems and Applications, Springer,* 13 (1), 1-19. https://doi.org/10.1186/s13677-023-00578-5

Kukkar, A., Shah, M. A., Mohana, R., Sharma, A. & Nayyar, A. (2023). Improving Sentiment Analysis in Social Media by Handling Lengthened Words. *IEEE Access*, 11, 9775 – 9788. **DOI:** 10.1109/ACCESS.2023.3238366

Li, Q., Peng, H., Li, J., Congyingxia, Renyuyang, Sun, L., Yu, P.S. & He, L. (2022). A Survey on Text Classification: From Traditional to Deep Learning. *ACM Transactions on Intelligent Systems and Technology*, 13 (2), 1-41. DOI:10.1145/3495162

Manias, G., Mavrogiorgou, A., Kiourtis, A. & Symvoulidis, C. (2023). Multilingual text categorization and sentiment analysis: a comparative analysis of the utilization of multilingual approaches for classifying twitter data. *Neural Computing and Applications, Springer*, 35, 21415–21431. https://doi.org/10.1007/s00521-023-08629-3

Md. Islam, S., Kabir, M.N., Ghani, N. A., Zamli, K.Z., Zulkifli, N. S. A., Md. Rahman, M., Moni, M.A. (2024). Challenges and future in deep learning for sentiment analysis: a comprehensive review and a proposed novel hybrid approach. *Artificial Intelligence Review, Springer*, 57 (62), 1-72. https://doi.org/10.1007/s10462-023-10651-9

Najafabadi, M. K. (2024). Sentiment analysis incorporating convolutional neural network into hidden Markov model. *Computational Intelligence, Wiley,* 2024, 1-28. DOI: 10.1111/coin.12633

Nedungadi, P., Raghuraman, Veena, G., Tang, K-Y. & Menon, R. K. (2025). AI Techniques and Applications for Online Social Networks and Media: Insights from BERTopic Modeling. *IEEE Access*, 13, 37389 – 37407**.** DOI: 10.1109/ACCESS.2025.3543795

Qiu, Q., Tian, M., Tao, L., Xie, Z. & Ma, K. (2024). Semantic information extraction and search of mineral exploration data using text mining and deep learning methods. *Ore Geology Reviews, Elsevier,* 165, 1-16. https://doi.org/10.1016/j.oregeorev.2023.105863

Rabby, G. & Berka, P. (2022). Multi-class classification of COVID-19 documents using machinelearning algorithms. *Journal of Intelligent Information Systems, Springer*, 60, 571–591. DOI: 10.1007/s10844-022-00768-8

Ranjan, N.M. & Chakkaravarthy, M. (2021). Evolutionary and Incremental Text Document Classifier using Deep Learning. *International Journal of Grid and Distributed Computing*, 14 (1), 587-595. https://www.lincolnedu.education/pdf/research/Evolutionary%20and%20Incremental%20Tex.pdf

Rivest, M., Vignola-Gagne, E. & Archambault, E. (2021). Article-level classification of scientific publications: A comparison of deep learning, direct citation and bibliographic coupling. *PLOSONE*, 16 (5), 1-18. https://doi.org/10.1371/journal.pone.0251493

Shyrokykh, K., Girnyk, M. & Dellmuth, L. (2023). Short text classification with machine learning in the social sciences: The case of climate change on Twitter. *PLOSONE,* 18 (9), 1-26. https://doi.org/10.1371/journal.pone.0290762

Taha, K., Yoo, P. D., Yeun, C., Homouz, D. & Taha, A. (2024). A comprehensive survey of text classification techniques and their research applications: Observational and experimental insights. *Computer Science Review, Elsevier*, 54, 1-21. https://doi.org/10.1016/j.cosrev.2024.100664

Tüselmann, O. & Fink, G.A. (2024). Neural modelsfor semantic analysis of handwritten document images. *International Journal on Document Analysis and Recognition, Springer*, 27, 245-263. DOI:10.1007/s10032-024-00477-8

Tyagi, S. & Szénási, S. (2024). Semantic speech analysis using machine learning and deep learning techniques: a comprehensive review. *Multimedia Tools and Applications, Springer,* 83, 73427–73456. https://doi.org/10.1007/s11042-023-17769-6

Uddin, K. M. M., Hamim, H., Mst. Mim, N. T., Akhter, A. & Md Uddin, A. (2024). Machine learning and deeplearning-based approach to categorize Bengali comments on social networks using

fused dataset. *PLOSONE*, 19 (10), 1-35. https://doi.org/10.1371/journal.pone.0308862

Wang, N-Q. (2024). Research on Deep Learning-Based Social Media Word-of-Mouth Analysis Model. *IEEE Access*, 12, 106537 – 106549. DOI: 10.1109/ACCESS.2024.3437734

Xia, B., Wang, X. & Yamasaki, T. (2021).Semantic Explanation for Deep Neural Networks Using Feature Interactions. *ACM Transactions on Multimedia Computing and Communication Applications*, 17 (115), 1-19. https://doi.org/10.1145/3474557

Yishun, Z., Guoyue, W., Yi, L., Yige, M. & Jiangwei, W. (2023). Classification of Distribution Network Planning Documents Based on LSTM Neural Network. *Procedia Computer Science, Elsevier*, 228, 914-919. https://doi.org/10.1016/j.procs.2023.11.120