AN HYBRID IMPLEMENTATION OF NLP AND TOPIC MODELLING IN THE TEXT CLASSIFICATION

Jiawei Zhang [1], Xin Zhang [2] and Xinyin Miao [3]

[1] Senior Investment Analyst, PRA Group (Nasdaq: PRAA), Norfolk, Virginia, USA
[2] Data Scientist, PRA Group (Nasdaq: PRAA), Norfolk, Virginia, USA
[3] Senior Data Analyst, American Airlines Group Inc (Nasdaq: AAL), Dallas, Texas

ABSTRACT

This article presents an innovative approach that combines quantified topic modelling results with TF-IDF based token features as a hybrid method for text classification. By integrating both techniques, the model is able to capture the contextual meaning of article text and improve overall classification performance. The hybrid approach quantifies the semantic meaningful words from article abstract, applies K-means clustering to group topics and then uses a gradient boosting model for the final classification task. Using five topics derived from a corpus of 750 articles, the proposed method improved the classification F1 score from 0.9121 to 0.9310 and accuracy score from 0.9115 to 0.9292. This approach offers a promising solution for long-form article classification scenarios, where topic modelling helps capture the core semantic meaning of paragraphs while complementing individual quantitative token features.

KEYWORDS

NLP, Text Classification, Topic Modelling, Machine Learning, Gradient Boosting

1. INTRODUCTION

With the most recent development of machine learning algorithms, natural language processing (NLP) has been gradually gaining more attention due to its ability to interact with human language and to quantify textual information for customer behaviour and preference prediction, on which ChatGPT and GenAI platforms heavily rely to build AI interaction platforms structure. Out of NLP implementation fields, text classification is an important area to divide the text input into corresponding categories to realize tasks ranging from classifying the textual inputs to supporting the backend RAG architecture to search for accurate content correlating to the input in LLM [1].

Traditional text classification approaches primarily rely on single-word quantification methods such as term frequency, document frequency, or TF-IDF, which often fail to capture the broader contextual meaning of text. While neural networks and transformer-based models can learn contextual relationships among words and sentences, they typically require large-scale datasets to achieve reliable performance. In data-limited scenarios, these deep learning models may underperform, whereas keyword-based methods remain insensitive to semantic context.

To address this gap, this study introduces a hybrid approach that leverages topic modeling through K-means clustering as a complementary feature extraction mechanism. By clustering documents into topic-based representations derived from TF-IDF features, K-means effectively captures high-level semantic similarity across texts. These topic cluster assignments are then

DOI: 10.5121/ijcsit.2026.18101

incorporated as additional quantitative inputs to the classification model, enabling contextual similarity awareness without the need for large training datasets.

2. LITERATURE REVIEW

NLP based classification machine learning models have been implemented and achieved promising accuracy in many fields. For example, NLP-based classification was implemented to categorize the academic curriculum related questions in Chatbots and achieved 98.7% accuracy using neural networks based on NLP processed data [2]. Classification using NLP techniques was also implemented in classifying the electronic health record (EHR) achieving 85% accuracy in EHR and 96% accuracy in NON EHR [3]. NLP based classification was also used for early detection of depression through quantifying the transcription of the conversation between a bot and a person as well as estimating the sentiment score of such conversation to classify the occurrence of depression through classification machine learning models (Random Forest & XGBoost), which achieves a 84% accuracy in depression early diagnosis [4]. Such techniques were also implemented in classifying the malware from benign groups through extracting ngrams, conducting feature selection and feature engineering to utilize the text information from API call sequences [5].

Topic modelling as an unsupervised machine learning model, was broadly researched in different scenarios. For example, BERT (Bidirectional Encoder Representations from Transformer) topic modelling was implemented in extracting the top 10 most commonly mentioned machine learning topics discussed in 45,783 industry 4.0 related articles [6]. Other topic modelling methodologies, such as LDA, have been implemented by Abderahman Rejeb to find 6 primary machine learning topics frequently discussed in 1114 publications regarding ML applications in agriculture fields [7]. When deciding the optimal number of topics through topic modelling, semantic coherence can also be used to find the number of topics that provides the highest semantic coherence score while separating the topics ideally, an comprehensive structural design used by Theresa Schmiedel to extract the 70 kinds of organizational cultures groups based on 428,492 employees reviews of Fortune 500 companies [8]. Sentiment analysis can also be complemented with the topic modelling to extract not only topics discussed in the article and reviews, but also the positive or negative sentiment of the comments. Such implementation was used in analyzing the COVID-19 vaccination related discussion on Twitter to tag the attitude of public regarding the vaccination [9].

To ensure the effectiveness of meaningful topics extraction, different methodologies can be used to conduct efficient topic modelling work. For instance, Bert-based Latent Semantic Analysis (LSA) gives higher U_mass scores and NPMI scores than Probabilistic Latent Semantic Analysis (PLSA) when evaluated based on 266 pieces of geospatial articles to find the topic of development trends over five years [10]. In another case study, LSA, LDA, NMF (Non-negative Matrix Factorization), PCA (Principal Dimension Analysis), and RP (Random Projection) were used in topic modelling for TF-IDF processed short-text data using F-score as evaluation metrics, with LDA outperforming the other topic modelling algorithms [11].

3. METHODOLOGY

The purpose of this article is to implement an innovative way to embed the topic modelling results as an additional input to improve the multi-class text classification prediction performance. Five classification target classes were selected as “Economics”, “Sports Science”, “Machine Learning”, “Geometry”, and “Art History”.

3.1. Data Extraction

To identify articles related to each target class, we used the academic article search functionality in Scopus [12], employing the five target classes as search keywords to retrieve relevant article abstracts. To further ensure a balanced topic distribution, we selected 150 abstracts from each target category, resulting in a total of 750 article abstracts and 153,812 words for training and testing purposes.

To compare the final classification prediction with and without the input of topic modelling for those 750 articles abstracts under 5 target classes, we've followed the processing workflow as shown in Figure 1 below:

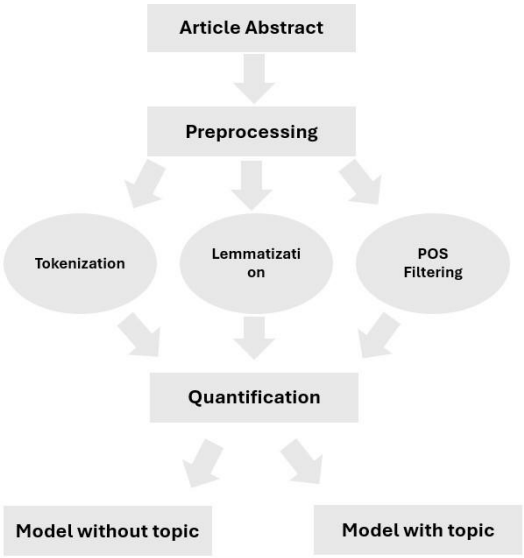

Figure 1. NLP and Classification Workflow

3.1. Preprocessing

In the preprocessing stage, we've conducted the three steps as below:

(1) **Tokenization:** The original article abstracts have been tokenized first to split the whole paragraph into word token for spacy word basis manipulation.
(2) **Lemmatization:** The tokens extracted from previous step are standardized by changing the plural, tense, and comparative forms into their word base roots using the ".lemma" method within spacy library. This step will make sure the meaning of different forms of the same word will be treated as the same word meaning input into the classification model.
(3) **POS filtering:**The final step is designed to filter out the Auxiliary, Conjunctions, Determiners and Numbers that don't carry the core meaning of the paragraph and to keep only the meaningful components such as Noun, Verb, and Adjectives.

The preprocessing adjustment keeps the core words of each abstract paragraph for more effective keywords, or numeric features after quantification, selection and remove the noise words from the paragraph. One of the sample paragraphs out of 750 paragraphs after NLP preprocessing is shown in Figure 2 below:

```
Abstract Before Preprocessing:

 This article reflects on a series of collaborations between artists in Chile, Maria Court and Trinidad Piriz, an interdisciplinary team of producers, an
d Matthew Brown, a UK historian of Latin America. We discuss the several manifestations of our collaboration over the four years of the estallido (social
upheaval) and its aftermaths, reflecting on the different creative approaches we took and how they enabled us to tell previously inaccessible stories. We
called it an 'essay about making history (unfinished)' not because we did not finish writing the essay, but because the history that was being made–a new
constitution, a new society–isn't over yet.

Abstract After Preprocessing:

 article reflect series collaboration artist Chile Maria Court Trinidad Piriz interdisciplinary team producer Matthew Brown UK historian Latin America di
scuss several manifestation collaboration year estallido social upheaval aftermath reflect different creative approach take enable tell previously inacce
ssible story call essay make history unfinished finish write essay history make new constitution new society yet
```

Figure 2. Abstract content before and after preprocessing

As shown in Figure 2, the preprocessing approach extracted the key meaning from the original text and successfully removed the high frequency but less meaningful words to avoid noises to be put into classification model. A more detailed token distribution can be found in Figure 3.

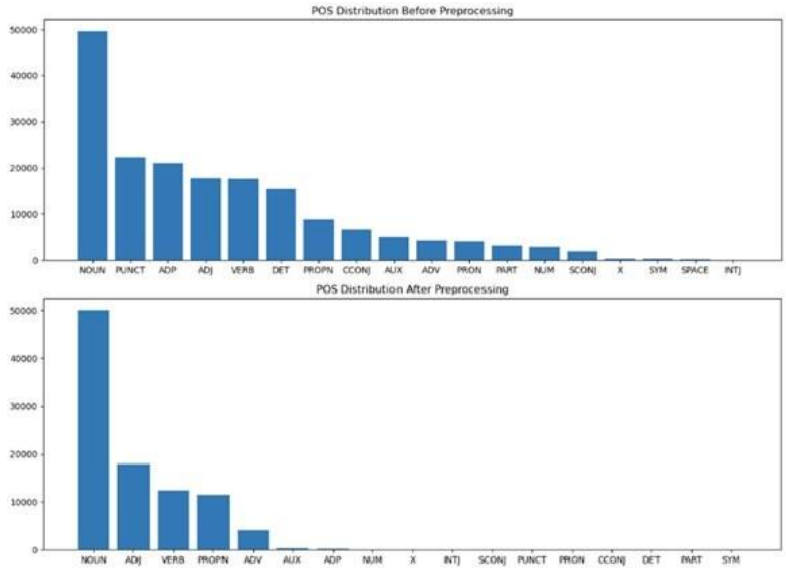

Figure 3. POS distribution before and after preprocessing

As shown in Figure 3, the preprocessing manipulation was able to tag and remove majority of Punctuations (PUNCT), Adpositions (ADP), Determiners (DET), Conjunctions (CCONJ), Auxiliaries (AUX) and other less meaningful words from more meaningful words.

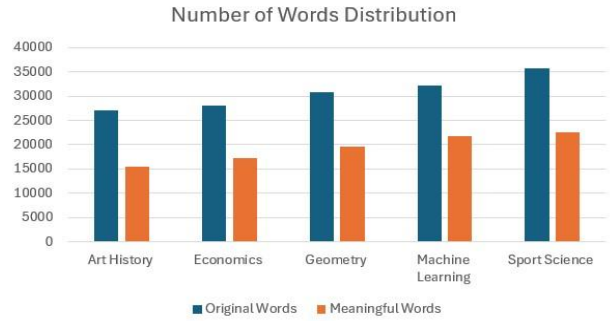

Figure 4. Number of words before and after preprocessing

Figure 4 shows the number of words distribution before and after texture preprocessing, of which the orange bars are the number of words that we were able to extract from the original abstracts to capture the main meaning. Due to the nature of academic topic, scientific classes have slightly more words in their abstracts than other classes.

The three texture preprocessing steps above was able to extract the 96,855 standardized word tokens from 153,812 tokens that carry core sentence meaning and are efficient to be used for further quantification.

3.2. Quantification

After the preprocessing, the selected tokens will be quantified into TF-IDF (Term Frequency – Inverse Document Frequency) score. TF-IDF score is the word representation methodology aimed at weighing the words that occur more frequently in individual documents (Term Frequency) but less frequently across different documents (Inverse Document Frequency), words that are beneficial to categorize and flag the document groups [13]. Such quantification will be beneficial for classification modelling to retrieve only the keywords that distinguish one target class from another.

The TF-IDF score transformation was conducted through TfidfVectorizer class from sklearn package. Two parameters are set up to extract as much meaningful information as possible.

(1) First, the ngram_range of TfidfVectorizer was set to be (1, 3), meaning that the TF-IDF transformer will quantify 1, 2, and 3 conjunctive words phrases from the processed documents to consider more context using longer phrases.
(2) Second, we've also set a limitation of min_df to be 0.15, which will constrain the quantification to be conducted for phrases that occurs in at least 15% of all 750 article abstracts, filtering out the rare nouns, such as a specific name or terminology, that occur highly frequently in only one or two articles but falsely end up with high TF-IDF score because of their high term frequency and low document frequency. A snapshot of quantified texture data is shown as in Figure 5.

	aim	also	analysis	analyze	approach	art	article	base	context	datum	...	sport	student	study	such	support	time	use
0	0.000000	0.000000	0.0	0.0	0.000000	0.887329	0.349703	0.0	0.000000	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000
1	0.000000	0.000000	0.0	0.0	0.291688	0.000000	0.000000	0.0	0.000000	0.0	...	0.0	0.0	0.391121	0.0	0.0	0.0	0.409986
2	0.000000	0.000000	0.0	0.0	0.204308	0.000000	0.234841	0.0	0.000000	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000
3	0.000000	0.107489	0.0	0.0	0.000000	0.575749	0.000000	0.0	0.126211	0.0	...	0.0	0.0	0.066175	0.0	0.0	0.0	0.000000
4	0.213479	0.000000	0.0	0.0	0.000000	0.471764	0.000000	0.0	0.000000	0.0	...	0.0	0.0	0.135557	0.0	0.0	0.0	0.000000

Figure 5. Processed texture data using quantified words

3.3. Topic Modelling

Besides the TF-IDF scores, this article also introduces topic modelling using K-means to cluster the documents into 5 topics to consider the relative term frequency among all selected phrases within each target class and thus introduce more context information beyond isolated phrase frequency.

The reason why we chose K-means as topic modelling approach is that K-means clustering for topic modelling has shown a better accuracy for keywords extracted NLP topic modelling when compared with other methodologies such as LDA. In the study [14], K-means model

outperforms LDA models in topic modelling by 5% WUP similarity between candidate labels and top 3 topic key words for journal and article text topic modelling.

Besides topic modelling accuracy, K-means as topic modelling approach also reduces the running time and shows a better distribution of words. In the study [15], when clustering for BBD news dataset, K-means methodology execution time under high number of topics is almost one-third of execution time of LDA and LSA, while still providing a higher CH-index than other topic modelling techniques.

3.4. Classification

Based on Subasish Das' analysis on comparison of classification performance among supportive vector machine, random forest, and Gradient Boosting, Gradient Boosting provides the best prediction performance in NLP classification for the studied textual data [16]. Therefore, in this article we use light gradient boosting model (Light GBM) to make the classification prediction.

3.5. Evaluation Metrics

To compare the prediction performance before and after topic modelling, we've introduced five testing classification metrics as below:

(1) Accuracy

Accuracy is the metric to measure how much percentage of testing target is correctly predicted by the classification model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100\%$$

(2) F1

F-1 score is the harmonic mean of precision and recall measuring the model's performance on positive prediction, especially if the target variable is imbalanced.

$$F1 = \frac{2 * TP}{(2 * TP + FP + FN)} * 100\%$$

(3) Recall

Recall is the metric to measure how much percentage of positive results are correctly labelled by the classification model as positive.

$$Recall = \frac{TP}{TP + FN} * 100\%$$

(4) Precision

Precision is the metric to measure how much percentage of positive predictions are actually true positive results.

$$Precision = \frac{TP}{TP + FP} * 100\%$$

(5) ROC
ROC_AUC (Receiver Operating Characteristic – Area Under the Curve) is the area under the ROC curve to represent the true positive rate against the false positive rate at various classification thresholds.

4. RESULTS

4.1. Topic Modelling Results

To mimic the actual implementation situation, the original 750 abstracts were randomly split into 85% training data and 15% validation data, of which the training data is used to train the K-means clustering model, and the rest 15% validation data is clustered by the model into corresponding topic clusters using the TF-IDF score of the 69 core keywords selected in the quantification process. After implementing the topic modelling using K-means, the 5 topics word distribution is shown as in Figure 6.

Topics & Keywords Distribution

	Sport Science	Sport Science Score	Machine Learning	Machine Learning Score	Geometry	Geometry Score	Economic	Economic Score	Art History	Art History Score
0	sport	0.557903	model	0.41439	student	0.178668	economic	0.574339	art	0.561161
1	study	0.142183	machine	0.201736	geometry	0.142321	social	0.097513	history	0.337216
2	research	0.119414	learning	0.151151	use	0.129065	research	0.083061	article	0.153448
3	student	0.110251	datum	0.131072	study	0.128724	study	0.082575	practice	0.082686
4	level	0.071546	use	0.125799	result	0.084762	article	0.077776	work	0.071612
5	practice	0.06717	study	0.104717	research	0.079941	approach	0.070642	new	0.0562
6	aim	0.065577	approach	0.090665	base	0.079685	paper	0.069889	paper	0.052656
7	use	0.06551	method	0.088333	design	0.075065	provide	0.069516	use	0.049865
8	education	0.064742	high	0.081349	education	0.069508	analysis	0.068215	focus	0.049401
9	result	0.064091	performance	0.078541	high	0.066749	also	0.060679	research	0.048954

Figure 6. Words distribution within each K-means topic

As shown in Figure 6, the keywords average TF-IDF score indicates the distribution of most representative keywords within each target class. For example, in the Machine Learning topic, highest average TF-IDF score of documents in this topic is 0.41439 for word “model”, with second and third highest score emphasize on word “machine” and “learning”. Therefore, in the testing scenario, article abstracts that share a similar score distribution as Machine Learning topic will be tagged as such topic accordingly.

After projecting the clustering results into 2 dimensions using PCA (Principal Component Analysis), the separation results of topic modelling are visualized as in Figure 7 as below. With 5 colors of dots representing the 5 topics, the K-means topic modelling shows a well-organized split pattern among different topics for those 750 article abstracts.

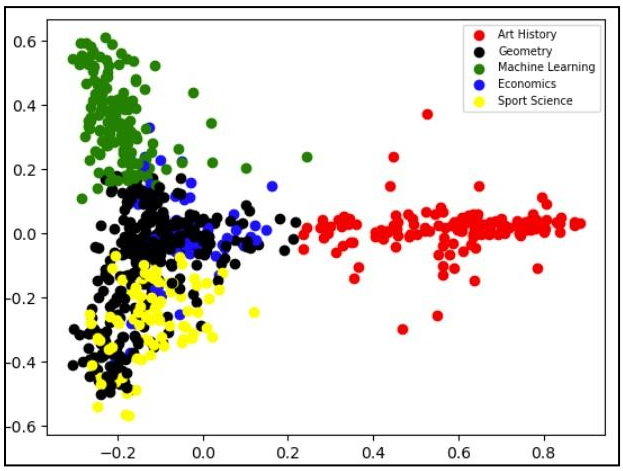

Figure 7. PCA distribution of K-means topic modelling

After implementing the preprocessing, quantification, and topic modelling, the final cleaned numeric data to be put into the classification model with and without topics can be found in Figure 8 and 5.

	target	topic	aim	also	analysis	analyze	approach	art	article	base	...	social	sport	student	study	such	support	time	use
0	Art History	Art History	0.000000	0.000000	0.0	0.0	0.000000	0.887329	0.349703	0.0	...	0.000000	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000
1	Art History	Geometry	0.000000	0.000000	0.0	0.0	0.291688	0.000000	0.000000	0.0	...	0.000000	0.0	0.0	0.391121	0.0	0.0	0.0	0.409986
2	Art History	Art History	0.000000	0.000000	0.0	0.0	0.204308	0.000000	0.234841	0.0	...	0.255329	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000
3	Art History	Art History	0.000000	0.107489	0.0	0.0	0.000000	0.575749	0.000000	0.0	...	0.246702	0.0	0.0	0.066175	0.0	0.0	0.0	0.000000
4	Art History	Art History	0.213479	0.000000	0.0	0.0	0.000000	0.471764	0.000000	0.0	...	0.000000	0.0	0.0	0.135557	0.0	0.0	0.0	0.000000

Figure 8. Processed texture data using K-means topics and quantified words

4.2. Classification Comparison Results

To eliminate the influence on performance comparison caused by subjective choice of prediction threshold, we've used the same threshold to examine the F1, recall, and precision performance for Light GBM with and without topic results as input. In addition, roc_auc score was also used as a comprehensive performance indicator that reflects the overall multi-class classification performance across all possible thresholds.

While the prediction F1 score using only TF-IDF phrase scores is 91.2%, the classification with topic modelling as additional feature has increased the F1 score to 93.1%. In addition, we've also compared the performance under other metrics as shown in Figure 9 below:

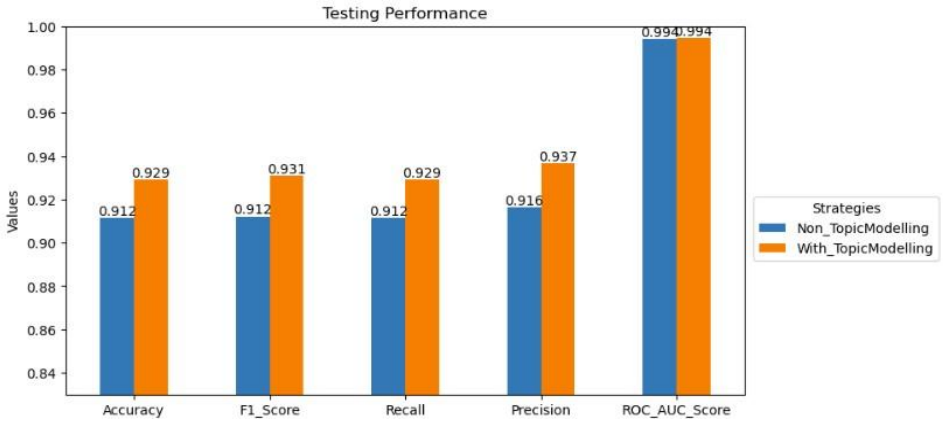

Figure 9. Testing performance comparison

Out of the 113 article abstracts in the testing data, 10 abstracts are wrongly predicted by using only TF-IDF score as features. When checking the false prediction in TF-IDF phrases model prediction, we've found that the K-means topic is able to cluster 4 of 10 false prediction results into the right topics and classification model using K-means topic as additional features corrected 3 out of 10 false prediction results as shown in Figure 10.

	Actual Target	Wrong Pred	Topic	Topic Adjusted Pred
1	Economics	Sport Science	Geometry	Geometry
2	**Economics**	Sport Science	Geometry	**Economics**
3	Geometry	Economics	Geometry	Economics
4	**Geometry**	Art History	Machine Learning	**Geometry**
5	Geometry	Sport Science	Geometry	Sport Science
6	Geometry	Economics	Geometry	Economics
7	**Machine Learning**	Geometry	Machine Learning	**Machine Learning**
8	Machine Learning	Economics	Geometry	Economics
9	Machine Learning	Geometry	Geometry	Geometry
10	Sport Science	Economics	Geometry	Economics

Figure 10. Topic modelling correction on false predictions

As shown in Figure 10, all 10 false prediction results made by TF-IDF based classification model are listed in the first two columns, of which the "Actual Target" column tags the actual result and the "Wrong Pred" column tags the wrong prediction of TF-IDF model. "Topic" column indicates the topics that are assigned to these 10 false predictions, of which 3 actual Geometry abstracts and 1 actual Machine Learning abstract are correctly captured by using topic modelling.

"Topic Adjusted Pred" column indicates the final prediction made by the classification model using both TF-IDF and topic as training features, which corrected 3 out of 10 abstracts that cannot be recognized by using TF-IDF based classification model.

5. CONCLUSIONS

This study demonstrates that incorporating topic modeling results alongside phrase-based TF-IDF features improves model performance and corrects misclassifications that are not captured by a TF-IDF based classification model. After adding topic modelling inputs, the accuracy and

F1 score for five-class text classification increased from 91.2% to 92.9% and 93.1% respectively, and approximately 30% of the misclassifications produced by the TF-IDF based model were corrected.

Despite these performance enhancements, the approach also has several spaces of improvement. The analysis is based on 750 article abstracts, and future research could expand the size of data to improve the generalizability of the results. In addition, this approach also shows the potential to be extended to less structured textual data with more complex features and targets for further NLP applications.

REFERENCES

[1] Muhammad Arslan, Hussam Ghanem, Saba Munawar, Christophe Cruz.(2024).A Survey on RAG with LLMs. Retrieved from: https://www.sciencedirect.com/science/article/pii/S1877050924021860

[2] Najma Rafifah Putri Syallya, Anindya Apriliyanti Pravitasari, and Afrida Helen. (2025).NLPBased Intent Classification Model for Academic Curriculum Chatbots in Universities Study Programs.Retrieved from https://jurnal.iaii.or.id/index.php/RESTI/article/view/6276

[3] K. Himavamshi, D. Tejaswini, Gaurav Sethi, V.S Anusuya Devi, P. Pavani, Shanmuga sundaram Hariharan. (2025). Electronic Health Record classification and analysis using NLP Techniques. Retrieved fromhttps://www.e3s-conferences.org/articles/e3sconf/pdf/2025/19/e3sconf_icsget2025_03016.pdf

[4] Giuliano Lorenzoni, Cristina Tavares, Nathalia Nascimento, Paulo Alencar, Donald Cowan. (2025).Assessing ML classification algorithms and NLP techniques for depression detection: An experimental case study. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC12118987/?tool=EBI#abstract1

[5] Bishwajit Prasad Gond, Rajneekant, Pushkar Kishore, and Durga Prasad Mohapatra. (2025).Malware Classification Leveraging NLP & Machine Learning for Enhanced Accuracy. Retrieved from https://arxiv.org/pdf/2506.16224

[6] Daniele Mazzei and Reshawn Ramjattan. (2022). Machine Learning for Industry 4.0: A Systematic Review Using Deep Learning-Based Topic Modelling. Retrieved fromhttps://mdpires.com/sensors/sensors-22-08641/article_deploy/sensors-22-08641.pdf?version=1667991055

[7] Abderahman Rejeb, Karim Rejeb, Abdo Hassoun. (2025).The impact of machine learning applications in agricultural supply chain: a topic modelling-based review. Retrieved from https://link.springer.com/content/pdf/10.1007/s44187-025-00419-1.pdf

[8] Theresa Schmiedel, Oliver Müller, and Jan vom Brocke. (2019).Topic Modelling as a Strategy of Inquiry in Organizational Research: A Tutorial With an Application Example on Organizational Culture. Retrieved from https://journals-sagepubcom.ezproxy1.lib.asu.edu/doi/epub/10.1177/1094428118773858

[9] Joanne Chen Lyu; Eileen Le Han; Garving K Luli. (2021).COVID-19 Vaccine–Related Discussion on Twitter: Topic Modelling and Sentiment Analysis. Retrieved from https://www.jmir.org/2021/6/e24435/pdf

[10] Quanying Cheng, Yunqiang Zhu, Jia Song, Hongyun Zeng, Shu Wang, Kai Sun and Jinqu Zhang. Bert-Based Latent Semantic Analysis (Bert-LSA): A Case Study on Geospatial Data Technology and Application Trend Analysis. (2021). Retrieved from: https://mdpires.com/applsci/applsci-11-11897/article_deploy/applsci-11-11897.pdf?version=1639494176

[11] Rania Albalawi, Tet Hin Yeap, Morad Benyoucef. Using Topic Modelling Methods for ShortText Data: A Comparative Analysis. Retrieved from: https://www.frontiersin.org/journals/artificialintelligence/articles/10.3389/frai.2020.00042/full#h1

[12] Scopus. Scopus Article Searching Functionality. Retrieved from: https://www.scopus.com/pages/home#basic

[13] Spärck Jones, K. (2004). IDF term weighting and IR research lessons, Journal of Documentation, Vol. 60 No. 5, pp. 521-523. https://doi.org/10.1108/00220410410560591

[14] Rahman, Shadikur ; Koana, Umme Ayman ; Ismael, Aras M. ;Abdalla, Karmand Hussein. (2025). Estimating the Effective Topics of Articles and journals Abstract Using LDA And KMeans Clustering Algorithm. Retrieved from: https://arxiv.org/pdf/2508.16046

[15] JUNAI D RASHID, SYED MUHAMMAD ADNAN SHAH, AUN IRTAZA. (2018). An Efficient Topic Modeling Approach for Text Mining and Information Retrieval through K-means Clustering. Retrieved from: https://www.researchgate.net/publication/338501601_An_Efficient_Topic_Modeling_Approach_for_Text_Mining_and_Information_Retrieval_through_K-means_Clustering

[16] Subasish Das, Anandi Dutta, Kakan Dey, Mohammad Jalayer, Abhisek Mudgal. (2020). Vehicle involvements in hydroplaning crashes: Applying interpretable machine learning. Retrieved from: https://www.sciencedirect.com/science/article/pii/S2590198220300877?pes=vor&utm_source=scopus&getft_integrator=scopus