KCMTI: A Framework for Knowledge Centric Microblog Tagging Integrating Incremental Knowledge Addition Paradigm and Quantitative Semantic Reasoning for Innovation and Strategy as a Domain of Choice

Anubrat Bora[1] and Gerard Deepak[2]

[1] Department of Computer Science and Engineering
Manipal Institute of Technology, Bangalore, Manipal Academy of
Higher Education, Manipal
[2] Department of Computer Science and Engineering
B.M.S Institute of Technology and Management, Yelahanka, Bengaluru
[2]gerard.deepak.christuni@gmail.com

Abstract. This paper presents a strategic and modern scheme for annotating microblogs, specifically targeting innovation and strategy as a domain of choice. In the era of Web 3.0, the proposed approach integrates semantic intelligence with robust machine learning and deep learning models within a unified architecture. This framework introduces an innovative and strategic knowledge stack, comprising eBooks, glossaries, abstracts, and keywords, which aids in generating high-density, domain-specific auxiliary knowledge for the model. The framework effectively shortlists informative terms for category extraction from microblog datasets, with knowledge enrichment achieved through metadata generation. This metadata is further inter-classified using a powerful RNN classifier. The top 10% of the classes identified by the RNN classifier are subsequently processed through an XGBoost classifier. Additionally, the model formalizes a semantic network and performs semantic similarity computations using Adaptive PMI (APMI) and the Horn's Index. This optimization enhances the model's overall precision, accuracy, and F-measure, achieving the lowest error rate, making it a best-in-class framework for knowledge-centric microblog tag.

Keywords: Semantic Intelligence, Adaptive PMI and Hans Index.

1 Introduction

In the rapidly evolving landscape of Web 3.0, the annotation of microblogs has become increasingly complex, particularly in specialized domains such as innovation and strategy. Addressing this challenge requires a sophisticated approach that integrates advanced semantic intelligence with robust machine learning and deep learning techniques. This paper introduces a comprehensive framework designed to enhance the annotation process for microblogs by leveraging a strategic and modern architecture. The proposed framework is anchored in the development of a rich knowledge stack that includes eBooks, glossaries, abstracts, and keywords, specifically tailored to the domain of innovation and strategy. This knowledge stack plays a crucial role in generating high-density, domain-specific auxiliary knowledge that informs the model. The process begins with the extraction of informative terms from microblog datasets, facilitated by the generation and classification of metadata. Initially, a Recurrent Neural Network (RNN) classifier is employed to perform metadata inter-classification. This deep learning model is adept at implicit feature selection and ensures that only the most relevant features are retained. The top 10% of the classified instances are then processed through an XGBoost classifier, a powerful ensemble learning method known for its accuracy and efficiency. This dual-classification approach significantly enhances the precision and relevance of the tagging process. Following classification, the framework formalizes a semantic network by computing semantic similarity using Adaptive Point-wise Mutual Information (APMI) and Horn's Index. These techniques optimize the model's performance, allowing for more precise and accurate tagging. APMI is employed to adjust the semantic similarity calculations, thereby improving the overall quality of the annotations. Horn's Index is used to set a significant level, ensuring that the model remains effective while accommodating the inherent diversity within the data. The integration of these advanced methodologies results in a framework that not only enhances the annotation of microblogs but also achieves superior performance metrics. The model demonstrates a significant reduction in error rates and an improvement in F-measure, positioning it as a leading solution for knowledge-centric microblog tagging. By combining semantic intelligence with cutting-edge machine learning techniques, this framework represents a significant advancement in microblog annotation.

.1 Motivation

The advent of Web 3.0 has underscored the necessity for frameworks that align with its principles, particularly in the context of tag recommendation for microblogs. The domain of innovation strategies has seen a significant surge in popularity, leading to an increased volume of microblogs. Despite this growth, a key challenge remains: the lack of semantic inclusion in existing tagging systems, even when additive knowledge schemes are utilized. This gap highlights the need for a standardized model that enhances the tagging process. The motivation behind this research is to develop a knowledge-centric framework for microblog tagging that integrates topic models with deep learning and machine learning techniques. By combining these approaches with quantitative semantic reasoning through semantic similarity measures, the proposed model aims to address the shortcomings in current systems and provide a robust solution for effective tag recommendation.

1.2 Contribution

The primary contributions of the proposed KCMTI framework are as follows: It introduces incremental knowledge addition by discovering informative terms within the dataset, enhancing the framework's ability to dynamically adapt and expand. The framework integrates a comprehensive knowledge stack that includes academic and research materials—such as e-books, glossaries, articles, abstracts, and keywords—within the innovation strategy domain. This innovative integration distinguishes the framework from existing approaches. By combining this knowledge stack with dataset-generated metadata and advanced topic models, the framework provides a robust tagging system. Additionally, it formalizes the semantic network by incorporating RNN-classified metadata and querying through SPARQL using an agent, which enhances mobility and flexibility. The dataset classification utilizes feedback from the top 10% of classes, inputted into the XGBoost Lightweight Measurement and Classifier, which is an effective approach. Furthermore, the framework employs Horn's Index with differential thresholds and centennial measures for meta-analytical optimization, resulting in a best-in-class integrative paradigm.

1.3 Organization

The paper's structure is as follows: A thorough list of books relevant to this paper is provided in Section 2. Section 3 presents an overview of the suggested system design, including the algorithm used to enhance the framework's performance. The outcomes produced by the proposed model are detailed in Section 4, with analysis of the results. Finally, Section 5 concludes the paper, discussing real-life implications of the model and its practical applications in various sectors.

2 Related Works

2.1 Semantic Annotation and Tagging

Shaw et al. [1] proposed Metablog: a semantics-aware, metadata-driven method for tagging blogs. Their framework leveraged metadata to enhance the semantic understanding and categorization of blog content, thereby improving tagging accuracy and relevance. Cassidy et al. (2012) [2] analyzed and enhanced wikification for microblogs by expanding context. Their approach improved the linking of microblog content to relevant Wikipedia entries, thereby enhancing the accuracy and relevance of the wikification process.

2.2 Opinion Summarization and Sentiment Analysis

Meng et al. [3] introduced an entity-centric, topic-oriented opinion summarization approach for Twitter. Their method focused on summarizing opinions around specific entities and topics by integrating topic modeling and sentiment analysis to produce concise opinion summaries. Fang et al. [4] addressed Word-of-Mouth Understanding through a multimodal, entity-centric opinion mining on social media. They proposed a model that integrated multiple modalities to analyze and extract aspects and opinions related to entities, thereby improving sentiment analysis and opinion mining. Li et al. [5] presented a method for mining opinion summarizations in Chinese microblogging systems using convolutional neural networks. Their approach leveraged deep learning techniques to extract and summarize opinions from microblogs, thus enhancing the quality of opinion analysis.

2.3 Content Retrieval and Search

Bandyopadhyay et al. [6] explored query expansion techniques for microblog retrieval. Their work aimed to enhance search results by expanding queries, addressing limitations in standard retrieval systems. Maniu et al. [7] investigated network-aware search in applications for social tagging. They examined the trade-offs between instance optimality and efficiency, focusing on effective search and retrieval of tagged content in social networks.

2.4 Topic Extraction and Community Detection

Li et al. [8] presented a method for topic extraction based on knowledge clusters in microblogs. Their approach utilized clustering techniques to improve the extraction of relevant topics from microblog data. Wang et al. [9] proposed a method for topic-aware interaction-centric detection and profiling of overlapping communities in microblogs. Their framework combined topic awareness with interaction-centric analysis to identify and profile overlapping communities within microblog platforms.

2.5 Mobile Systems and Geolocation in Microblogging

Gaonkar et al. [10] explored the integration of mobile technology and social participation in microblogging platforms. Their work investigated how mobile systems facilitate the sharing and querying of content, emphasizing the role of mobile phones in content dissemination and retrieval within social networks. Their study highlighted the impact of mobile technology on enhancing user engagement and accessibility in microblogging. In a complementary vein, Di Rocco et al. [11] introduced Sherloc, a knowledge-driven algorithm designed for precise geolocation of microblog messages at the sub-city level. Their method improved location inference by integrating geographic and contextual information, which significantly enhanced the accuracy of geotagging within urban areas.

3 Methodology

This section embeds the proposed methodology and architecture for the Knowledge Centric Microblog Tagging and Quantitative Semantic Reasoning for Innovation and Strategy as a domain of choice.

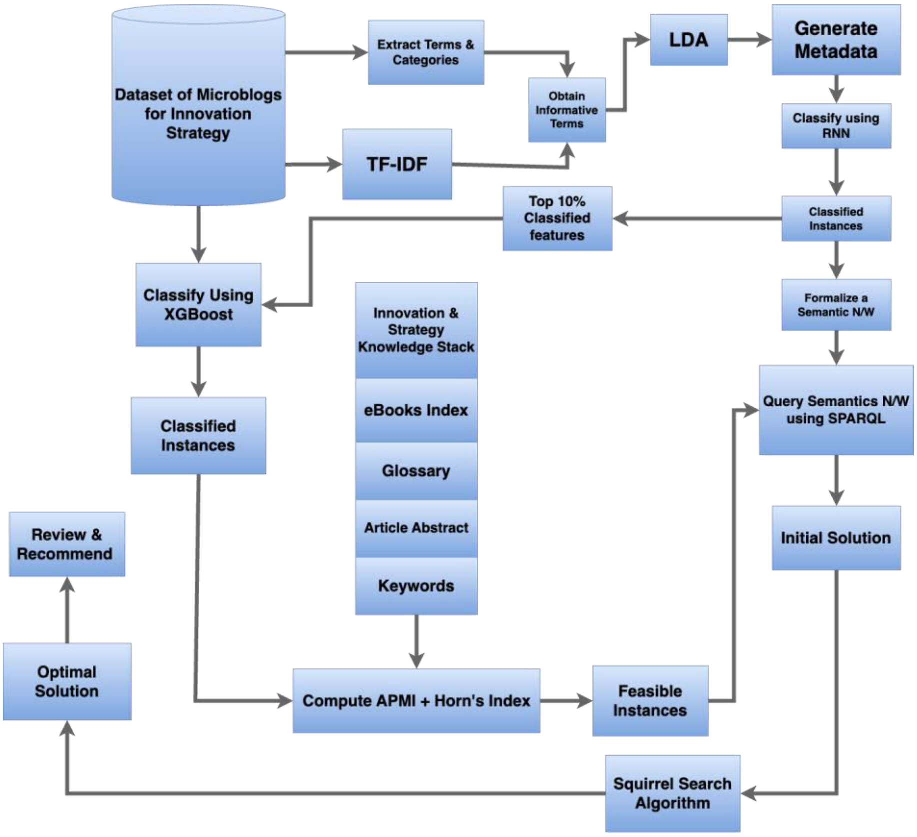

Fig. 1. Proposed Architecture

Fig. 1. demonstrates the proposed system architecture for a strategic framework designed for microblog tagging, with innovation and strategy as the chosen domain of focus. The framework begins with the dataset of microblogs related to innovation strategy, which undergoes classification using the XGBoost algorithm. This algorithm processes and classifies the metadata derived from the dataset. The microblogs are subjected to term and category extraction, where terms are directly obtained from the dataset's keywords, and categories are sourced from predefined categories within the dataset. To shortlist and obtain informative terms, the Term Frequency-Inverse Document Frequency (TF-IDF) is calculated across the document corpus. These informative terms are then analyzed using Latent Dirichlet Allocation (LDA), a topic modeling technique.

Given the extensive size of the dataset, initial classification is necessary before further processing. The dataset is first classified using a Recurrent Neural Network (RNN), a robust deep learning classifier that performs implicit feature selection. The RNN classifies the metadata from the dataset itself. Subsequently, only the top 10% of the classified instances are selected as features and fed into the XGBoost classifier to further classify the microblogs dataset. This selection of features is sufficient for effectively classifying the microblogs within the innovation and strategy domain. The classified dataset is processed through an ensemble of classifiers, including XGBoost. An innovation strategy knowledge stack is then formalized, consisting of eBooks, indexes, glossaries, articles, abstracts, and keywords. The articles, abstracts, and keywords are drawn from research papers, while glossaries and indexes from relevant eBooks are combined to create a comprehensive knowledge stack, which is subsequently visualized. The knowledge stack visualization provides classified outputs based on adaptive point-wise visual information. Hans' index is set to a significance level of 0.15, and the Adaptive Point-wise Mutual Information (APMI) is set to a threshold of 0.75. The relaxation of Hans' index is justified due to the strength of APMI, which has been empirically adjusted to enhance diversity and increase the number of feasible instances. These feasible instances are utilized within the framework. Classifier instances generated by the random classifier are further processed to formalize a semantic network by computing the information measure of the nodal terms using Shannon's entropy. After review by domain experts, the drop-down solution is finalized and used to recommend tags or annotations, which are then applied to the relevant categories and labels directly within the dataset.

Algorithm 1: Knowledge-Centric Microblog Tagging Framework for Innovation and Strategy

```
  1. Load Dataset:
     - Load microblogs dataset related to "innovation strategy"

  2. Classification:
     - Use XGBoost to classify microblogs based on metadata

  3. Term and Category Extraction:
     - Extract terms and categories from dataset keywords

  4. TF-IDF and LDA:
     - Calculate TF-IDF scores for informative terms
     - Apply LDA for topic modeling on selected terms

  5. Initial Classification with RNN:
     - Train RNN on metadata for initial classification
     - Select top 10% instances for further processing

  6. Further Classification with XGBoost:
     - Classify top instances using XGBoost for detailed categorization

  7. Incremental Knowledge Addition:
     - Continuously update knowledge stack with new research data and glossaries

  8. Quantitative Semantic Reasoning:
     - Use APMI and Hans' Index to enhance semantic diversity and instance feasibility
```

```
9. Visualization:
  - Visualize classified outputs with adaptive point-wise visual information

10. Information Measure:
  - Compute Shannon's entropy for semantic network analysis

11. Tagging and Annotation:
  - Finalize tags and annotations with expert review
  - Apply tags to relevant categories and labels in the dataset
```

Algorithm 1. utilizes a knowledge-centric approach to tag microblogs focused on innovation and strategy. It employs pseudocode-like steps to load and classify data using XGBoost, extract terms with TF-IDF and LDA, and refine classifications via RNN and additional XGBoost. Continuous updates and quantitative reasoning enhance semantic diversity. Visualizations and Shannon's entropy analysis aid comprehension, while expert-reviewed tags and annotations are applied systematically within the dataset.

3.1 Term Frequency-Inverse Document Frequency (TF-IDF)

The relevance of a term within a document in relation to a corpus is determined using the Term Frequency-Inverse Document Frequency (TF-IDF) equation. It combines two essential metrics: inverse document frequency (IDF), which quantifies how common or uncommon a term is over the entire corpus, and term frequency (TF), which counts how frequently a term shows up in a document. Mathematically, it is depicted as shown in equation 1.

$$TF - IDF = TF \times \log \frac{N}{DF} \tag{1}$$

where N is the total number of documents, and DF is the number of documents containing the term. The TF-IDF score is higher for terms that appear frequently in a document but are rare across the corpus, making it useful for identifying important words in context.

3.2 Recurrent Neural Network (RNN) Classifier

A Recurrent Neural Network (RNN) is a type of deep learning model that is particularly effective for sequential data, such as time series or text. Because feedback loops are a part of the RNN's architecture, it may remember and utilize information from earlier calculations. An RNN's parameter settings usually consist of the number of epochs, batch size, learning rate, and hidden layers. The model's capacity to learn from and generalize the data is influenced by these factors. An RNN can be efficiently tailored to classify complicated datasets, like microblogs within areas, by fine-tuning these settings.

3.3 XGBoost Classifier

The gradient boosting framework is powerfully and effectively implemented by XGBoost (Extreme Gradient Boosting). It functions by constructing a group of decision trees, each of which aims to fix the mistakes of the ones before it. Because of its speed and scalability, XGBoost is a popular choice for handling big datasets. Regularization to avoid overfitting, parallel processing to expedite computation, and sparse awareness to manage missing values are important components. Because of XGBoost's excellent accuracy and performance, it is widely used in machine learning competitions as well as real-world applications.

3.4 Google Search Algorithm

An advanced algorithm called the Google Search Algorithm is used to get the most pertinent results for a user's query. It analyzes and ranks web pages using a variety of ranking signals and algorithms. Important parts are Hummingbird, which concentrates on comprehending the context and intent behind a query, and PageRank, which ranks the value of web pages according to the quantity and caliber of links. Furthermore, by using data from previous searches, RankBrain, a machine learning component, assists Google in processing and interpreting difficult questions. The algorithm is modified frequently to increase relevance and accuracy.

3.5 Hans Index

The Hans Index is a tool for determining how important phrases or features are in a dataset. It offers a mechanism to rank features for additional examination or categorization by assisting in determining the significance of phrases based on their frequency and distribution. Depending on the precise implementation and environment in which it is employed, the particulars and computation techniques may change.

3.6 APMI (Adaptive Point-wise Mutual Information)

Adaptive Point-wise Mutual Information is referred to as APMI. This sophisticated metric evaluates the relationship between terms in natural language processing and information retrieval. By modifying the mutual information depending on observed frequencies, APMI adjusts to the frequency and distribution of terms in a dataset, providing a more nuanced understanding of term interactions. As a result, it can be helpful for determining important phrases and enhancing feature selection accuracy.

4 Results

The experimentations utilized six distinct datasets: Government of Canada Statistics Canada (2014), focusing on innovation and business strategy, including process innovation and reduction of average unit costs [15]; Jiayu Liu et al. (2023), providing measures of textual innovation [16]; GlobalData UK Ltd. (2024), covering the innovation and patenting activity of Nippon Paper Industries Co. Ltd. in Q2 2024 [17];

Technavio (2021), offering market insights into document outsourcing [18]; World Bank (2022), analyzing global innovation metrics and economic impact [19]; and Gartner (2023), providing market analysis on emerging technologies and innovation trends [20]. These datasets underwent remediation with common annotations and keywords and were integrated into a comprehensive dataset based on these annotations and categories. A common annotator further refined the annotations, while fraudulent documents were identified and merged into a single dataset, resolving all discrepancies. The KCMTI framework was effectively applied to these real-world datasets, extracting key themes and trends. For example, it analyzed the Government of Canada Statistics Canada (2014) data to provide insights into national innovation strategies and examined the GlobalData UK Ltd. Q2 2024 dataset to reveal trends in patenting activity and innovation strategies for Nippon Paper Industries Co. Ltd. These case studies illustrate KCMTI's practical utility in analyzing innovation trends in both public and corporate sectors. Future work should include conducting experiments on larger datasets to evaluate the framework's scalability and provide insights into its computational efficiency and performance in real-time.

Table 1. Comparision of the Proposed KCMTI with other approaches.

Model	Average Precision %	Average Recall %	Average Accuracy %	Average F-Measure %	FDR
SABT [12]	90.22	93.45	91.84	91.84	0.10
ASTA [13]	90.74	93.87	92.31	92.31	0.10
DMTS [14]	90.86	94.01	92.44	92.44	0.10
Proposed KCMTI	94.74	96.16	95.45	95.45	0.06

From Table 1, it is evident that the proposed KCMTI model outperforms every other basic model in terms of precision and accuracy. The KCMTI demonstrates superior performance, achieving the highest metrics in these areas. In contrast, the other baseline models show lower performance, which can be attributed to their less advanced methodologies compared to the KCMTI. The KCMTI's enhanced capabilities and integration of advanced techniques contribute to its superior results. The performance of the proposed KCMTI model, designed for microblog tagging with an incremental knowledge addition strategy and quantitative semantic reasoning tailored for innovation strategies, is assessed using various metrics. The primary metric is the precision-indicator accuracy, reflecting the model's ability to accurately tag microblog, while the secondary metric involves the sterilization of FDRIs (False Discovery Rate Indices). The proposed KCMTI demonstrates exceptional performance, with the highest average precision of 94.74%, the highest average recall of 96.16%, and the highest average accuracy and F-measure among the compared models. In evaluating and comparing the proposed KCMTI, baseline performance is established using four distinct 3D-streamed analysis frameworks, including ASTA, SABT and DMTS. The final step involves analyzing the results to determine the model's effectiveness relative to the baseline frameworks. This paper introduces a novel strategic framework for annotating microblogs within the domain of innovation strategies, specifically tailored for the era of Web 3.0. It integrates advanced semantic intelligence techniques with robust machine learning and deep learning models within a unified architecture. This approach leverages cutting-edge technology to enhance the accuracy and relevance of microblog annotations, addressing the unique challenges of the innovation strategies domain.

The KCMTI is a highly domain-specific, knowledge-centric paradigm designed for microblog tracking, with a particular focus on innovation and strategy. The model stands out from baseline models due to its strategic integration of a knowledge stack that includes e-books, indexes, glossaries, article abstracts, and keywords. This stack provides high-density, domain-specific background knowledge that significantly enhances the framework's performance. Additionally, the KCMTI framework derives knowledge from a dataset perspective by employing specialized strategies to obtain informative terms. Techniques such as term frequency and adjusted frequency measures contribute to the framework's ability to identify and utilize significant terms effectively. Moreover, the KCMTI model is integrated with topic modeling frameworks like Latent Dirichlet Allocation (LDA) and leverages metadata generation from the Web 3.0 perspective. This integration involves utilizing the structural metadata from Web 3.0 and classifying it using a robust Recurrent Neural Network (RNN) deep learning classifier. This approach enhances the formalization and accuracy of the semantic network, yielding well-defined classified instances. Additionally, feedback from the auxiliary classified metadata, which includes the top 10% of the classes, is used as features for the XGBoost classifier, a lightweight yet effective machine learning model. This process ensures precise classification of the dataset. Furthermore, semantic similarity is computed using adaptive point-wise mutual information measures and Horn's Index at various stages of the pipeline. These measures, adjusted with differential thresholds, help in assessing deviations and refining the overall classification process. This model employs a joint multi-label attention network, utilizing bidirectional units and hierarchical attention mechanisms. While the learning mechanism is robust, the underlying knowledge of the metadata is relatively shallow. This shallow knowledge causes underfitting, in which case the model may not fully capture and incorporate the best-in-class terms. Despite the strong learning mechanism, this limitation affects the model’s ability to leverage high-quality terms effectively.

Let’s examine the SABT model. The SABT model is designed for semantic annotation of microblog topics using temporal and contextual information. While temporal and contextual information provides valuable auxiliary knowledge and supports semantic similarity discovery, the density of the augmented auxiliary knowledge is somewhat limited. Consequently, the learning paradigm within this model does not fully address this limitation, leading to less effective performance compared to the proposed framework. Similarly, the DMTS model lags in its ability to effectively integrate and utilize available data. While it operates on a multi-labeled corpus of Twitter short texts and employs a semi-automatic method, it does not fully leverage temporal and contextual information. This limitation impacts its overall performance, especially when compared to more advanced models like KCMTI that integrate sophisticated techniques such as deep learning classifiers, hybrid approaches, and robust feedback mechanisms. As a result, the DMTS model may struggle with classification accuracy and the effective handling of complex datasets. Our model, KCMTI, leverages a sophisticated hybrid classification approach that sets it apart from existing models. It employs informative term derivation through strategies such as term frequency-inverse document frequency (TF-IDF) and topic modeling using Latent Dirichlet Allocation (LDA). Metadata generation and classification are executed with a robust Recurrent Neural Network (RNN) deep learning classifier. Feedback from the top 10% of the classified metadata, obtained from the RNN, is used as input for the XGBoost classifier to further refine the dataset classification. This combination of techniques ensures superior performance in classifying and annotating microblog data. KCMTI's integration of RNN and XGBoost classifiers, along with precise feedback mechanisms, significantly enhances accuracy and effectiveness compared to existing models. By incorporating TF-IDF, LDA, and a hybrid classification approach, KCMTI effectively handles the complexity of large datasets and achieves robust, efficient classification results.

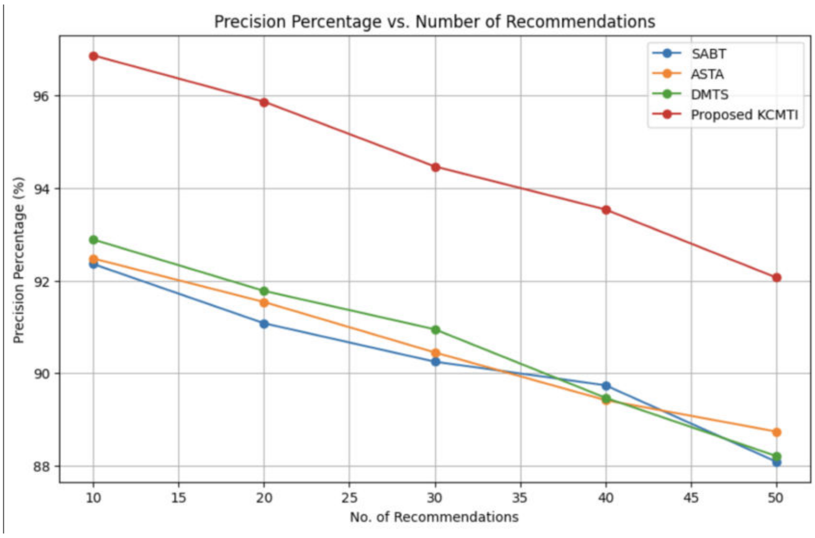

Fig. 2. Graph showing Precision Percentage vs. Number of Recommendations

As depicted in **Fig. 2**, the hierarchical positioning of the baseline models illustrates the performance of KCMTI at the top of the hierarchy. The SABT model occupies the lowest portion, ASTA is positioned just above SABT, and DMTS is situated second from the top. KCMTI's prominence in the hierarchy is due to its comprehensive approach, which integrates advanced techniques such as TF-IDF, LDA, and a robust hybrid classification system involving both RNN and XGBoost classifiers. This results in a substantial performance advantage, reflected in its higher precision and accuracy. In contrast, SABT, ASTA, and DMTS lag for several reasons. SABT's limitations include restricted auxiliary knowledge density and lack of effective temporal information utilization. ASTA's performance is hampered by its less advanced feature integration and classification techniques. DMTS, while semi-automatic, does not fully leverage temporal and contextual data, resulting in less effective classification. Consequently, KCMTI's superior approach and integration of advanced methodologies explain its higher ranking in the hierarchy.

5 Conclusion

The paper introduces the KCMTI framework, a knowledge-centric approach for microblog tagging that integrates incremental knowledge addition with quantitative semantic reasoning. Utilizing a logic modeling-based LDA approach, the framework incrementally discovers and enriches informative terms, generating diverse metadata that is managed by an RNN-based strategy classifier. Focused on innovation strategy, KCMTI incorporates various resources such as articles, abstracts, keywords, e-books, indices, and glossaries, enabling it to handle the evolving nature of innovation-related microblogs effectively. The framework's performance in FHWR planning highlights its robustness, with the theorem-based system showing the lowest value. Its real-world effectiveness is demonstrated through accurate tagging of datasets from government and corporate sectors, providing valuable insights into national policy strategies and corporate patenting activities.

References

1. Shaw, Harsh, and Gerard Deepak. "MetaBlog: a metadata driven semantics aware approach for blog tagging." *International Conference on Digital Technologies and Applications*. Cham: Springer International Publishing, 2022.
2. Cassidy, Taylor, et al. "Analysis and enhancement of wikification for microblogs with context expansion." *Proceedings of COLING 2012*. 2012.
3. Meng, Xinfan, et al. "Entity-centric topic-oriented opinion summarization in twitter." *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2012.
4. Fang, Quan, et al. "Word-of-mouth understanding: Entity-centric multimodal aspect-opinion mining in social media." *IEEE Transactions on Multimedia* 17.12 (2015): 2281-2296.
5. Li, Qiudan, et al. "Mining opinion summarizations using convolutional neural networks in Chinese microblogging systems." *Knowledge-based systems* 107 (2016): 289-300.
6. Bandyopadhyay, Ayan, et al. "Query expansion for microblog retrieval." *International Journal of Web Science* 1.4 (2012): 368-380.
7. Maniu, Silviu, and Bogdan Cautis. "Network-aware search in social tagging applications: Instance optimality versus efficiency." *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 2013.
8. Li, Ming, et al. "Topic extraction based on knowledge cluster in the field of micro-blog." *Intelligent Computing Methodologies: 10th International Conference, ICIC 2014, Taiyuan, China, August 3-6, 2014. Proceedings 10*. Springer International Publishing, 2014.

9. Wang, Zhu, et al. "Topic-aware Interaction-centric Overlapping Community Detection and Profiling in the Microblog." *Journal of Multiple-Valued Logic & Soft Computing* 31 (2018).
10. Gaonkar, Shravan, et al. "Micro-blog: sharing and querying content through mobile phones and social participation." *Proceedings of the 6th international conference on Mobile systems, applications, and services*. 2008.
11. Di Rocco, Laura, et al. "Sherloc: a knowledge-driven algorithm for geolocating microblog messages at sub-city level." *International Journal of Geographical Information Science* 35.1 (2021): 84-115.
12. Tran, Tuan, et al. "Semantic annotation for microblog topics using wikipedia temporal information." *arXiv preprint arXiv:1701.03939* (2017).
13. H. Dong, W. Wang, K. Huang and F. Coenen, "Automated Social Text Annotation With Joint Multilabel Attention Networks," in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 5, pp. 2224-2238, May 2021, doi: 10.1109/TNNLS.2020.3002798.
14. Liu, X.; Zhou, G.; Kong, M.; Yin, Z.; Li, X.; Yin, L.; Zheng, W. Developing Multi-Labelled Corpus of Twitter Short Texts: A Semi-Automatic Method. *Systems* **2023**, *11*, 390. https://doi.org/10.3390/systems11080390
15. Government of Canada, Statistics Canada (2014). Innovation and business strategy, process innovation introduced; reduction of average unit cost of existing products [Dataset]. http://doi.org/10.25318/2710005101-eng
16. Jiayu Liu; William E Underwood; Sarah Griebel; Lucian Li; Rebecca Cohen; Jana M. Perkins; Jaihyun Park (2023). Comparing Measures of Textual Innovation [Dataset]. http://doi.org/10.17605/OSF.IO/A3G6E
17. GlobalData UK Ltd. (2024). Innovation and Patenting activity of Nippon Paper Industries Co Ltd Q2 2024 [Dataset]. https://www.globaldata.com/store/report/innovation-and-patenting-activity-of-nippon-paper-industries-co-ltd-innovation-and-trend-analysis/
18. Technavio (2021). Document Outsourcing Market Insights [Dataset]. https://statistics.technavio.org/statistics/document-outsourcing-market-insights
19. World Bank. (2022). *Global Innovation Metrics and Economic Impact*. [Dataset]. https://data.worldbank.org/
20. Gartner. (2023). *Market Analysis on Emerging Technologies and Innovation Trends* [Dataset]. https://www.gartner.com/