Literature Survey on Application and Development Trend of n-gram Model in NLP

CS6120 Natural Language Processing SEC 01 Summer full 2024

Zhanshuo Dai
Northeastern University
dai.zha@northeastern.edu

Hiranmai Devarasetty
Northeastern University
devarasetty.h@northeastern.edu

Krishna Priya Gitalaxmi
Northeastern University
gitalaxmi.k@northeastern.edu

Snehitha Barukula
Northeastern University
barukula.s@northeastern.edu

***Abstract*—Influential development in NLP using n-gram models has made huge differences in improving a variety of applications and solving many prevailing issues. This survey highlights the development and application trends on the n-gram models to enhance sentiment analysis, language prediction, and data efficiency. This survey aims to consider hybrid approaches in which the n-gram model is fused with other machine learning methods in details so as to increase classification accuracy and effectiveness. Finally, we will consider tools developed for the treatment of large n-gram datasets, language modeling for minority languages, and prospects for the integration of n-gram models in quantum computing. The survey highlights the versatility of n-gram models in practical applications by covering new uses of the n-grams in text classification for chatbots, predictive text systems, and even malware detection.The survey offers insights into the effectiveness of n-gram models and suggests future research areas to overcome persistent issues in NLP by using extensive data from Google Scholar.**

I. Introduction

The rapid progress in NLP has overaccentuated the requirement of an efficient language model and its accuracy. Among them, N-gram models have continued to remain in circulation only because of their ease and efficiency in tasks such as text prediction, sentiment analysis, and language translation. Though one of the oldest techniques in NLP, an n-gram model continues to evolve with modern machine learning methods to address current challenges. The ability of these models to capture the statistical properties of word sequences makes them very instrumental in many NLP tasks. This literature survey explores trends in applying and developing n-gram models, outlining their relevance for the field of Machine Learning.

A recent effort in this direction is represented by a survey that studies the current state of the art regarding applications and development of n-gram models in NLP. It demonstrates how n-gram models are used in combination with other techniques of machine learning to boost performance and innovative applications in various domains—polarization in sentiment analysis of social media platforms and, indeed, advanced topics such as malware detection and integrations in quantum computing. Such future research directions will be brought to light by identifying the actual challenges and limitations that n-gram faces in modern NLP tasks.

The criteria that had to be used in the selection of papers for this survey were quite strict to obtain a comprehensive and up-to-date review. The selected research papers were published within the last five years, thus reporting the most current developments and trends. Papers selected for the review were derived from databases indexed on Google Scholar, assuring their quality and relevance. Publications in high-impact factor journals and conferences were given priority, indicative of rigorous peer-review processes and significant contributions. It included works that deal explicitly with the application, enhancement, or innovative use of n-gram models in NLP to maintain a focused and coherent survey.

The integration of n-gram models with other machine learning techniques has led to notable advancements. For instance, hybrid models combining n-grams with Naïve Bayes have shown improved accuracy in sentiment classification tasks on social media data [1]. Similarly, the introduction of ProphetNet, which utilizes future n-gram prediction for sequence-to-sequence tasks, demonstrates how traditional n-gram approaches can be adapted for more sophisticated models [2]. These examples highlight the versatility and adaptability of n-gram models in enhancing performance and addressing complex NLP challenges. Furthermore, innovative applications such as the dynamic n-gram system for minority languages and efficient handling of massive n-gram datasets reflect ongoing efforts to optimize and expand the applicability of these models [3], [4].

In addition to these advancements, the survey reviews significant contributions such as the use of n-gram models for enhancing text prediction and entertainment systems [5], and the integration of convolutional and LSTM layers in cloud computing environments for improved sentiment classification [6]. The development of NLP models for Hindi, addressing the challenges of text autocompletion [7], and the replication of human contextual behavior using web mining techniques for decision-making [8] are also examined. By utilizing comprehensive data from Google Scholar, this survey provides insights into the effectiveness of n-gram models and identifies future research directions to address ongoing challenges in NLP.

II. Methodology

A. Search Strategy

We have sourced a large amount of relevant literature in regard to the application and development trends of n-gram models in NLP from multiple databases and search engines. The primary resources included IEEE Xplore and Google Scholar, both of which are renowned for their extensive collections of peer-reviewed papers and articles in the field of computer science and machine learning. By leveraging these platforms, we were able to identify key papers and articles that discuss the evolution, implementation, and future directions of n-gram models in NLP.

B. Keywords

The search process was guided by a carefully chosen set of keywords and phrases to ensure the retrieval of relevant literature. The primary keywords used in the search included ”n-gram model”, ”Natural Language Processing”, ”sentiment analysis”, ”quantum computing”, ”federated learning”, ”text prediction”, ”machine learning”, ”language model”, ”text classification”, ”chatbot”, ”malware detection”,”genomics” and ”file compression.” These terms were combined in various configurations to maximize the breadth and depth of the search results.

C. Selection Criteria

Ensuring that the literature is relevant and of high quality, a set of strict inclusion and exclusion criteria was applied while selecting papers. First, on relevance: only works that explicitly addressed the application or development of n-gram models within the domain of NLP were considered. This focus ensured that all studies included contributed directly to the understanding and further development of n-gram methodologies in NLP contexts. The publication period was limited to the last five years, from 2019 to 2024, setting a base of recent developments and trends in the area. This allowed this review to portray the most current research, technologies, and methodologies being investigated and put into practice within the domain of n-gram models. Other key factors in the selection process were quality and impact. Papers targeting high-impact journals and conferences have been included since these sources reflect rigorous peer review and hence important contributions to the field. High-impact publications are normally subjected to intensive vetting, placing a high degree of reliability and scientific merit on the studies published. Not least important, citation count has been used as an added filter for including publications. Papers with a greater number of citations were more desirable, as frequent citations normally indicate that the paper is influential and widely recognized by the research community. In other words, the highest citation counts show that the work, by and large, has been accepted and useful; therefore, it is of prime importance and relevance for modern academic discourses. These criteria, taken together, have meant that the selected literature will present a very strong overview of high quality regarding the current state and recent developments in the application of n-gram modeling to natural language processing and, hence, support a comprehensive, in-depth literature survey. In summary, 20 papers were selected based on this criterion and they will be analyzed in this article.

III. Literature Review

A. Thematic Analysis

N-gram Models in Sentiment Analysis and Text Classification
N-gram models are one of the cornerstones in both sentiment analysis and text classification. Their ability to capture and utilize such contextual information makes them indispensable tools for understanding natural language. Among the 20 selected articles, 7 of them, that is, most of the articles, applied n-gram to text classification, which shows the effectiveness of n-gram in this task.

Recently, research has begun that expands the application of n-gram and tries to combine it with other algorithms to achieve better results. In this respect, Awwalu et al. [1] discussed the use of a hybrid n-gram model combined with Naïve Bayes for the purpose of classifying political sentiment on Twitter. This study showed improved accuracy by leveraging n-grams, which helped capture contextual sentiment nuances in social media posts. Similarly, Liu et al. [9] introduced a hierarchical attention architecture that utilizes n-grams in conjunction with CNN and Bi-LSTM layers to detect spam reviews. Kotiyal et al. [10] explored the use of n-grams for text classification to enhance chatbot responses, showing significant improvements in user interaction and satisfaction. Rajesh and Suseendran [11] applied n-gram models in sentiment analysis of e-learning reviews, enhancing the understanding of student feedback and course effectiveness. Furthermore, Ghude et al. [7] contributed an NLP model for Hindi text generation that has made the interaction of the user more plausible in applications such as chatbots and content-based tools. These studies underline the effectiveness of n-grams in improving performance in sentiment analysis and text classification tasks. An emerging approach that enhances these capabilities is federated learning. [12] This machine learning technique that allows multiple decentralised devices to train models collaboratively without the need of sharing raw data. This approach significantly enhances the privacy, making it particularly relevant for applications handling sensitive data, such as tweet classification. By performing model training on devices,federated learning ensures that user data remains secure and private,thus addressing a critical concerns in modern NLP applications.

Advanced Applications of N-gram Models
The integration of n-gram models with deep machine learning techniques enhanced many applications in the domain of NLP. The bar graph categorizes the number of papers based on different advanced applications of N-gram models. Most of the papers (10) focus on information technology, highlighting its prominence in the application of N-gram models. Healthcare/bioinformatics and educational technology also have significant research interests. Cybersecurity, while

an emerging field, has a smaller number of papers, but is still notable.

The line chart on the right shows the trend of research papers and hottest topics over the five-year period from 2018 to 2023. The number of research papers has steadily increased from around 6 in 2018 to 9 in 2023. There are different “hottest topics” each year, indicating that the research focus areas have diversified over time.

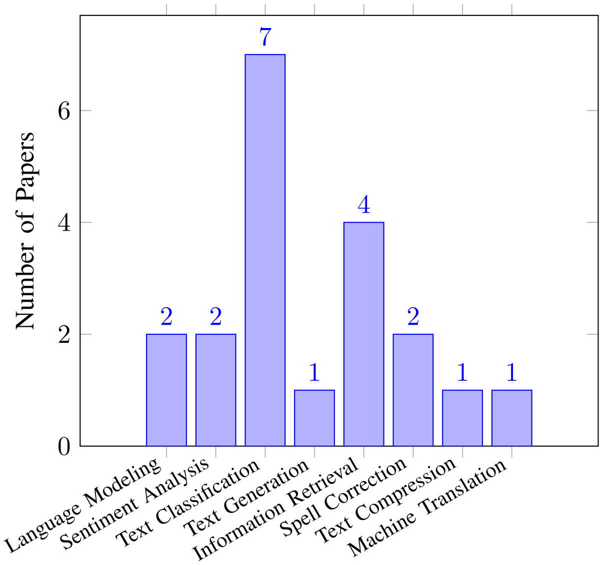

Fig. 1. Applications of N-gram Model across papers

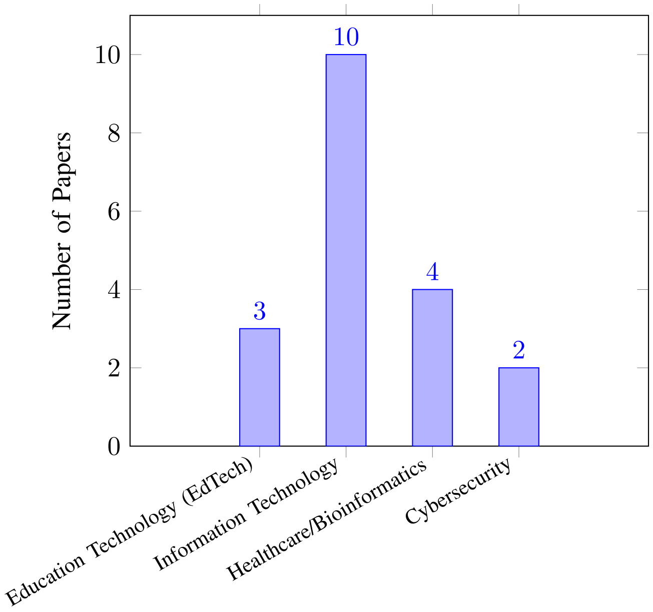

Fig. 2. Thematic Analysis of N-gram Model Research

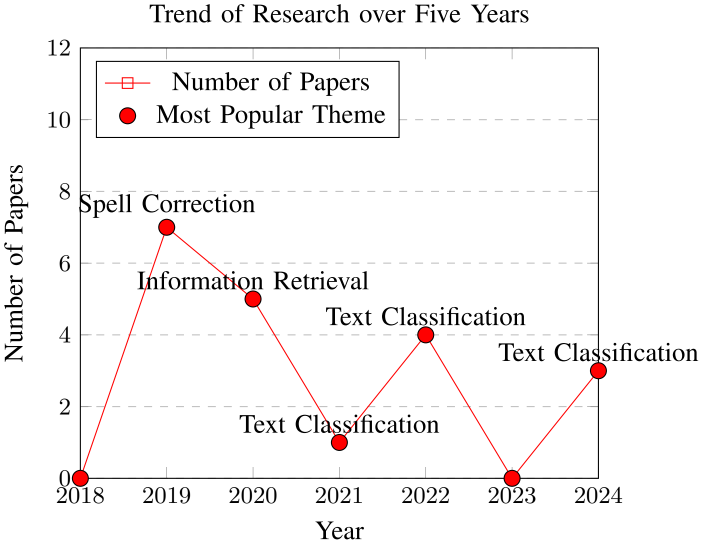

Fig. 3. Research Papers and Most Popular Themes over Five Years

Research into n-grams has featured in the generation and translation of minority languages. Gledec et al. [3] discussed some of the problems in the development and maintenance of large-scale n-gram systems for minority languages. They contribute a dynamic n-gram system that was part of an online Croatian spell-checking service, describing practical challenges and solutions to the creation of robust language models for less spoken languages. Ghude et al. [7] complement this by providing an NLP model for Hindi and enhancing user experience in many applications, from chatbots to content-based tools. Further, Gowda et al. [8] introduced a new sentence n-gram model that can be used for predicting user actions on social media comments. It is able to combine dynamic user profiling with the real-time intake of data and machine learning techniques to construct improved prediction accuracy with ensemble methods. Binggui Zhou et al. [13] also reviewed the significant progress in smart healthcare enabled by emerging AI technologies.This review covers NLP applications and approaches in clinical health,public health, personal health, hospital management and drug development. This paper addresses the role of NLP in addressing the COVID-19 pandemic and mental health issues, highlighting the current limitations and future directions. John S.Malamon’s study introduces the DNA N-gram Analysis Framework (DNAnamer) designed for the supervised classification of DNA sequences. [14].This framework applies N-gram methodologies to genomics, facilitating tasks such as gene identification, species classification, and contamination detection. The study highlights the conserved and identifiable nature of DNA N-gram frequency patterns, demonstrating their applicability across various genomic contexts.

N-gram models have been applied in several critical domains such as healthcare and cybersecurity. Yazdani et al. [15] applied the trigram model for word prediction in electronic health records to improve the efficiency of documentation. This model cut down a number on typing time and keystrokes. The finding is thus very promising toward the improvement of EHR systems. Dou et al. [16] checked the application of an enhanced n-gram algorithm in malware detection, proving that the n-gram models can handle semantic analysis for assembly codes in identifying unknown threats. Islam et al. [17] put forward an N-gram-based multi-layer approach for the detection of Android malware and assesses different techniques for N-gram to increase malware detection accuracy and efficiency. Marellapudi [18] showed how n-grams can be efficiently used in lossless file compression to create smaller files without any loss of data. Agarwal and Keselj [19] showed the usage of n-grams for mental health disorder detection in social media, demonstrating how the model can perform in health-related text classification.

Efficient Handling of Massive N-gram Datasets

An important requirement for the scalability of NLP applications is how to efficiently process and store large n-gram datasets. More than half of the articles try to answer how to process vast data volumes. Pibiri and Venturini [4] proposed a compressed data structure that efficiently handled huge n-gram datasets, compressing them greatly while maintaining high retrieval speed. It is crucial in applications like search engines and machine translation that require fast and effective data processing. Kotiyal et al. [10] have added that it is also essential in chatbot performance.

B. Comparative Analysis

Methodologies

The methodologies used across these works differ vastly, ranging from applications and challenges in NLP. From the perspective of the algorithms used, all papers have their own innovative points, but we also found some common algorithms, which we put in Table 1. The hybrid approach with n-grams combined with Naïve Bayes [1] stands far away from deep learning methods like a hierarchical attention network [9]. While the former resorts to probabilistic models for enhanced sentiment classification, the latter integrates convolutional and LSTM layers for the capturing of multi-granularity semantic information. Kotiyal et al. [10] implemented n-grams in text classification to improve chatbot responses by integrating n-grams with machine learning algorithms, demonstrating significant enhancements in user interaction quality. Rajesh and Suseendran [11] applied n-grams for sentiment analysis in e-learning reviews, using a combination of feature selection methods and machine learning techniques to better understand student feedback. Marellapudi [18] utilized n-grams for lossless file compression, focusing on redundancy in text data to achieve efficient compression. Dou et al. [16] explored an improved n-gram algorithm for unknown malware detection, combining n-grams with advanced machine learning techniques to identify malicious patterns in software. Agarwal and Keselj [19] applied the Common N-Gram (CNG) method to detect mental health disorders on social media, leveraging n-grams to identify linguistic markers indicative of mental health issues. Ghude et al. [7] applied a probabilistic n-gram approach for text generation in Hindi, while Gowda et al. [8] used a combination of n-gram models with machine learning techniques in user behavior prediction. Islam et al. [17] presented a multi-layer approach to improve malware detection. These methodological differences underline what the approach of n-gram models is capable of in being adapted to various tasks of NLP. John S.Malamon [14] applied higher n-grams to genomics,playing a vital role in classifications tasks.

TABLE I
FREQUENCY OF METHODS USED IN N-GRAM MODEL RESEARCH PAPERS

Method	Count of Papers
Naïve Bayes	5
SVM	3
Hidden Markov Models (HMM)	2
Logistic Regression	2
k-Nearest Neighbors (k-NN)	2
Deep Learning	2
Decision Trees	1
Huffman Coding	1
Run-Length Encoding (RLE)	1
Random Forest	1
Gradient Boosting	1
Deep Neural Networks (DNN)	1
Convolution Neural Network (CNN)	1
Hierarchical Attention Network	1
Minimal Perfect Hash (MPH)	1
Kneser-Ney Smoothing	1
Markov Chain Model	1
Recurrent Neural Network (RNN)	1

Datasets

The datasets used in these studies also differ, ranging from social media data [1] to clinical text data [15]. The choice of dataset influences the model design and evaluation, with some studies focusing on language-specific challenges (e.g., Hindi NLP models by Ghude et al. [7]) while others address more generalizable issues (e.g., handling massive n-gram datasets by Pibiri & Venturini [4]). For instance, Kotiyal et al. [10] utilized chatbot interaction logs, Rajesh and Suseendran [11] analyzed e-learning reviews to enhance sentiment analysis models, Marellapudi [18] focused on English text data for lossless file compression, Dou et al. [16] worked with software code for unknown malware detection, and Agarwal and Keselj [19] used social media posts for mental health detection. Hamarashid et al. [5] presented a comprehensive study on predictive and entertainment text-based applications, covering datasets like MovieLens for the evaluation of recommendation systems and Yelp for sentiment analysis. Ghorbani et al. [6] proposed a ConvLSTMConv network for accomplishing sentiment analysis in cloud computing and trained and evaluated it using the MR2004 movie reviews dataset. These differences underline how dataset selection is very instrumental in shaping the scope and applicability of the n-gram models.

Results and Conclusions

The results and conclusions across these papers demonstrate the strengths and limitations of n-gram models. For instance, the improved accuracy in sentiment classification achieved by hybrid models [1], [11] and the efficient handling of large datasets [4], [18], Dou et al. [16] showcased effective malware detection using improved n-gram algorithms, Agarwal and

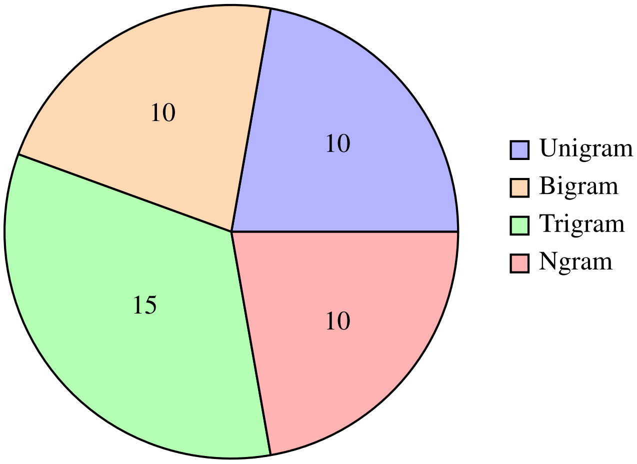

Fig. 4. Usage of Unigram, Bigram, and Trigram Models in Research Papers

Keselj [19] successfully applied n-grams to detect mental health disorders on social media.These showcase the practical benefits of n-grams. Hamarashid et al. [5] reviewed text predictive systems, underpinning the strength of n-grams in entertainment applications. Ghorbani et al. [6] demonstrated the use of n-grams in cloud computing in sentiment analysis, thus their application domains significantly vary across different technological domains. The computational efficiency of quantum n-gram models [20], the scalability of n-gram systems [3] for minority languages raise some challenges which should be targeted by further research and development.

Strengths and Weaknesses

The greatest strengths of the n-gram models lie in their simplicity and the ability to encompass some form of context about the text. This makes them very efficient at tasks such as sentiment analysis and text prediction. However, the models still have major weaknesses, such as strong dependence on historical data and difficulty in handling out-of-vocabulary words. The comparative analysis indicated that, while the n-gram models were quite effective in some contexts, advanced machine learning techniques and domain-specific adaptations were important integrations necessary to overcome limitations and expand their scope of applications.

IV. Critical Analysis

A. Evaluation of Research Quality

The studies reviewed in this literature survey demonstrate a range of methodologies and results, each contributing uniquely to the field of n-gram models in NLP. Most of these methods differ in their validity and reliability since some studies are very rigorous in terms of their experimental design, with complete datasets, while others suffer from the limitations of sample size or less strong methodologies. For instance, Awwalu et al. [1] provided a thorough validation of their hybrid n-gram model using Naïve Bayes through extensive testing on Twitter data, showcasing significant improvements in sentiment classification accuracy. Similarly, Liu et al. [9] employed a hierarchical attention architecture with CNN and Bi-LSTM layers, demonstrating robust performance in spam review detection.

It is equally compelling because of the practical applications and improvements that present the importance of the research. Qi et al. [2] introduced ProphetNet, a new way to do future n-gram prediction, having implications for improving language models in many NLP tasks. In another approach, Ghorbani et al. [6] utilized models that process information in sequences using ConvLSTMConv in handling sentiment analytics over cloud-based computing resulting in 87% accuracy and 86% precision. The integration of quantum computing with n-gram models by Payares et al. [20] highlights a cutting-edge advancement that could revolutionize computational efficiency and performance in NLP applications. These studies have been conducted to build on continuously evolving n-gram models, which are relevant and applicative in modern NLP research.

TABLE II
Evaluation Metrics and Observed Values for N-gram Model Studies

Study	Accuracy	Precision	Recall
Sentiment Analysis E-Reviews [11]	80%	80%	80%
Multi-Class Mental Health Disorder [19]	NA	27%	31%
Unknown Malware Detection [16]	100%	100%	100%
Text Generation in Hindi [7]	69%	70.2%	69.7%
User Behavior Prediction [8]	98%	NA	NA
Malware Detection in Android [17]	88%	87%	86%
Sentiment Analysis in Cloud Computing [6]	89.02%	NA	NA
Classification of Political Sentiments [1]	89%	NA	NA
Detection of Spam Reviews [9]	83%	86.8%	85.3%

TABLE III
Evaluation Metrics and Observed Values for N-gram Model Studies contd.

Study	F1-Score
Sentiment Analysis E-Reviews [11]	80%
Multi-Class Mental Health Disorder [19]	20%
Unknown Malware Detection [16]	100%
Text Generation in Hindi [7]	69.9%
User Behavior Prediction [8]	NA
Malware Detection in Android [17]	86%
Sentiment Analysis in Cloud Computing [6]	NA
Classification of Political Sentiments [1]	80%
Detection of Spam Reviews [9]	86.1%

B. Identification of Gaps

Despite the significant advancements, several gaps in the current research have been identified. One notable gap is the limited focus on minority languages, as highlighted by Gledec et al. [3]. While their work on Croatian n-gram models is commendable, there remains a need for more extensive research on other less commonly spoken languages. Additionally, the scalability and efficiency of n-gram models in handling massive datasets, as discussed by Pibiri and Venturini [4], indicate potential areas for further optimization, particularly in terms of reducing computational overhead and improving retrieval speeds.

Another gap is the integration of n-gram models with emerging technologies such as quantum computing. While Payares et al. [20] have made strides in this area, further research is needed to fully explore the potential and limitations of quantum-enhanced n-gram models. Additionally, the application of n-gram models in diverse fields such as healthcare and cybersecurity, as explored by Yazdani et al. [15] and Dou et al. [16], respectively, suggests that interdisciplinary research could yield innovative solutions and broaden the impact of n-gram models. Islam et al. [17] demonstrated that multi-layer neural networks were actually quite efficient for malware detection; however, the possibility of using this approach against other strains or types of malware, or other security threats, is not fully researched.

Kotiyal et al. [10] focused on immediate user interaction quality in chatbot responses, suggesting a need for exploring long-term user engagement and more advanced conversational AI techniques. Rajesh and Suseendran's [11] work on sentiment analysis in e-learning reviews was limited to specific datasets, indicating the potential for broader studies across diverse platforms and languages. Marellapudi [18] demonstrated efficient lossless file compression mainly on English text data, pointing to the need for testing on multilingual datasets and different data types like images and videos. Agarwal and Keselj's [19] research on detecting mental health disorders on social media was limited by the scope of platforms analyzed, highlighting the need for studies across various social media networks and more extensive linguistic and cultural contexts to enhance the robustness and applicability of their findings.

C. Implications

The findings from the reviewed studies have both practical and theoretical implications. On a practical note, improvements in n-gram models may lead to the development of more accurate, faster NLP applications, such as better sentiment analysis tools, predictive text systems, and spam and malware detection mechanisms. For instance, the trigram model for word prediction in EHR by Yazdani et al. [15] has the potential to significantly enhance healthcare documentation efficiency, thereby improving clinical workflows and patient care.

In theory, this would further unify models of n-grams with advanced machine learning techniques, for instance, the hierarchical attention network from Liu et al. [9], and the exploration of quantum computing by Payares et al. [20], contribute to the broader understanding of how traditional NLP models can be augmented and improved. These theoretical insights will help to ensure that, against the changing backdrop of technology, future research in n-gram model optimization and fine-tuning is able to further develop and fine-tune such models so as to retain their relevance and effectiveness.

D. Limitations

On the other hand, those selected studies also have a number of limitations that have to be acknowledged.

Looking at Fig. 5, we also find some patterns. Scalability and adaptability are the most frequently mentioned limitations, appearing in 13 research papers. This shows that many studies face challenges when trying to scale their solutions or adapt them to different environments. Platform scope is the second most common limitation, appearing 8 times. This shows that the range of platforms on which research solutions can operate effectively is often limited.

One limitation is the differences in the type and size of datasets used, which might have an effect on the generalizability of the results. For example, studies performed on social media data [1] do not apply directly to other text domains, such as clinical or legal documents. In addition, methods relying on historical data for training of their n-gram models have little hope of performing well for words out of their vocabulary and when linguistic trends change rapidly.

The literature review itself has limitations, including a bias towards more recent studies and those published in high-impact journals. That approach bears a number of limitations, particularly to the effect that it may include only high-quality pieces of research, leaving useful insights from older studies or less prominent publications. More importantly, the specific focus on applications like sentiment analysis and text prediction using n-gram models restricts how comprehensive the review would be when representing all possible uses and developments in n-gram model research.

While these studies that have been reviewed help greatly in compiling information on the trends and developments of n-gram models in NLP, an appreciation of the limitations brought out in this study is important for proper balance in view. In such a view, future research should be aimed at addressing these gaps and limitations for further development of the field and widening of its applications through the use of n-gram models in diverse, emerging contexts within the NLP domain.

V. Conclusion

A. Summary of Findings

An in-depth literature review on the applicability and development trends of n-gram models in NLP underlines the fact that such models are timeless and versatile. The overall outlook from these reviewed works has been able to demonstrate just how n-gram models have effectively been combined with various machine learning techniques for improved performance in sentiment analysis, text classification, and language prediction, besides enhancing data efficiency. Major ones include the hybrid n-gram and Naïve Bayes model for political sentiment analysis over Twitter [1], hierarchical attention architecture incorporating CNN and Bi-LSTM for spam review detection [9], and future n-gram prediction model ProphetNet [2]. Other innovative applications of n-gram models include the use of quantum computing for classifying tweets [20] and developing n-gram models for languages spoken by much smaller populations, such as Croatian and Hindi [3], [7].

This can be further exemplified by the works of Pibiri and Venturini, who have tried to deal with huge n-gram datasets

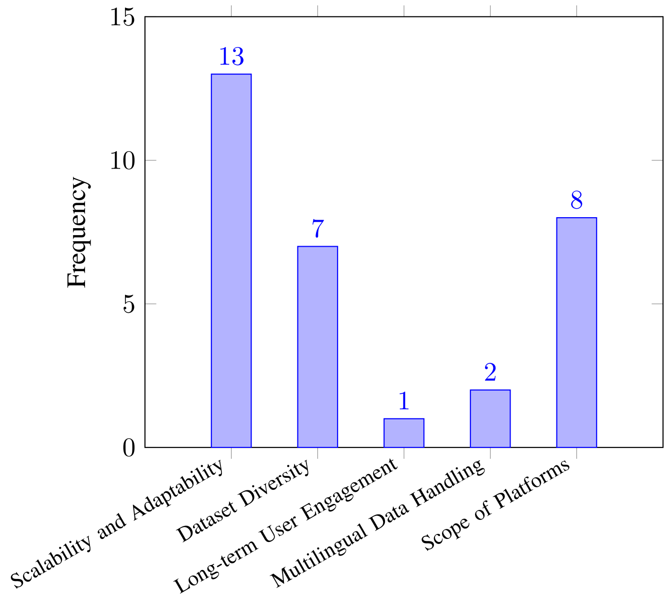

Fig. 5. Frequency of Limitation Categories in Research Papers

effectively [4], and the models' application in health care as well as in cyber security [15], [16] brings out more broad application and practical benefits of n-grams. However, gaps were identified through this literature review as it relates to the further optimization of handling big data, in-depth research into minority languages, and quantum-enhanced n-gram models.

B. Future Directions

With respect to the identified gaps and emerging trends, several areas for future research can be suggested. First, the absence of more comprehensive studies about the application of n-gram models for minority languages is apparent; this could comprise developing methodologies that are more robust for languages with limited digital resources or include ensuring that NLP improvements will benefit other linguistic communities. The second critical area remains the optimization of the scalability and efficiency of n-gram models in handling large datasets. One could do research on new compression techniques or distributed computing environments to manage and process huge n-gram datasets more efficiently.

On the other hand, another promising line of research is the integration of n-gram models with emerging technologies, would be quantum computing. While initial studies like those by Payares et al. [20] have shown potential, further research is needed to fully understand the capabilities and limitations of quantum-enhanced n-gram models. Moreover, interdisciplinary research on the application of models of n-grams in domains like healthcare and cybersecurity needs to be broadened. Each of these applications will be driven by a development of domain-specific adaptations and optimizations related to n-gram models.

Finally, handling out-of-vocabulary words and extremely dynamic linguistic trends may include a high degree of improvement in the robustness of n-gram models. Hybrid approaches, merging n-grams with neural network-based approaches, might provide greater flexibility toward the changing language trends in use in the future. Based on these lines, research in the future will be able to continue making steps forward in the area of N-gram models within NLP, in order to keep them relevant and effective in the continuously changing technological landscape.

References

[1] J. Awwalu, A. Abu Bakar, and M. R. Yaakub, "Hybrid n-gram model using naïve bayes for classification of political sentiments on twitter neural computing and applications," *Neural Computing and Applications*, 2019.

[2] W. Qi, Y. Yan, Y. Gong, D. Liu, N. Duan, J. Chen, R. Zhang, and M. Zhou, "Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training," *arXiv preprint arXiv:2001.04063*, 2020.

[3] G. Gledec, R. Šoić, and Dembitz, "Dynamic n-gram system based on an online croatian spell checking service," in *2019 IEEE*, 2019.

[4] G. E. Pibiri and R. Venturini, "Handling massive n-gram datasets efficiently," *ACM Transactions on Information Systems*, 2019.

[5] H. K. Hamarashid, S. A. Saeed, and T. A. Rashid, "A comprehensive review and evaluation on text predictive and entertainment systems," *Soft Computing*, 2022. [Online]. Available: https://doi.org/10.1007/s00500-021-06691-4

[6] M. Ghorbani, M. Bahaghighat, Q. Xin, and et al., "Convlstmconv network: a deep learning approach for sentiment analysis in cloud computing," *Journal of Cloud Computing*, vol. 9, no. 16, 2020. [Online]. Available: https://doi.org/10.1186/s13677-020-00162-1

[7] T. Ghude, R. Chauhan, K. Dahake, A. Bhosale, and T. Ghorpade, "N-gram models for text generation in hindi language," in *ITM Web of Conferences*, vol. 44, no. 03062, 2022, pp. 1–5.

[8] B. N. S. Gowda and V. Lakshmikantha, "User behavior prediction using a novel sentence n-gram model," in *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, 2020, pp. 391–397.

[9] Y. Liu, L. Wang, T. Shi, and J. Li, "Detection of spam reviews through a hierarchical attention architecture with n-gram cnn and bi-lstm," *arXiv preprint arXiv:2105.01761*, 2021.

[10] A. Kotiyal, P. J. Gujjar, G. P. M. S., and P. H. R. Kumar, "Text classification using n-grams for providing effective response in chatbot," in *2023 International Conference on Computer Science and Emerging Technologies (CSET)*, 2023, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/10346678

[11] P. Rajesh and G. Suseendran, "Prediction of n-gram language models using sentiment analysis on e-learning reviews," in *2020 International Conference on Intelligent Engineering and Management (ICIEM)*, 2020, pp. 510–514. [Online]. Available: https://ieeexplore.ieee.org/document/9160260

[12] M. Chan, A. T. Suresh, R. Mathews, A. Wong, C. Allauzen, F. Beaufays, and M. Riley, "Federated learning of n-gram language models," in *23rd Conference of Computational Natural Language Learning (CoNLL)*, 2019. [Online]. Available: https://arxiv.org/abs/1910.03432

[13] B. Zhou, G. Yang, Z. Shi, and S. Ma, "Natural language processing for smart healthcare," *IEEE Reviews in Biomedical Engineering*, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9904944

[14] J. S. Malamon, "Dna n-gram analysis framework (dnanamer): A generalized n-gram frequency analysis framework for the supervised classification of dna sequences," *bioRxiv*, 2024. [Online]. Available: https://doi.org/10.1101/2024.02.02.578674

[15] A. Yazdani, R. Safdari, A. Golkar, and S. R. N. Kalhori, "Words prediction based on n-gram model for free-text entry in electronic health records," *Health Information Science and Systems*, vol. 7, no. 6, 2019. [Online]. Available: https://doi.org/10.1007/s13755-019-0065-5

[16] J. Dou, Y. Zhang, Y. Shi, and L. Jiang, "Improved n-gram algorithm for unknown malware detection," in *2022 International Conference on Machine Learning, Cloud Computing and Intelligent Mining (MLCCIM)*, 2022, pp. 165–170. [Online]. Available: https://ieeexplore.ieee.org/document/9955154

[17] T. Islam, S. S. M. M. Rahman, M. A. Hasan, A. S. M. M. Rahaman, and M. I. Jabiullah, “Evaluation of n-gram based multi-layer approach to detect malware in android,” *Procedia Computer Science*, vol. 171, pp. 1074–1082, 2020. [Online]. Available: https://doi.org/10.1016/j.procs.2020.04.115

[18] H. Marellapudi, “Lossless file compression using redundant ngrams in english,” in *2019 Global Conference for Advancement in Technology (GCAT)*, 2019, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/8978441

[19] H. Agarwal and V. Keselj, “Common n-gram method (cng): A promising approach to detecting mental health disorders on social media,” in *23rd International Symposium INFOTEH-JAHORINA*, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10495942

[20] E. Payares, E. Puertas, and J. C. Martinez-Santos, “Quantum n-gram language models for tweet classification,” in *2023 IEEE 5th International Conference on Cognitive Machine Intelligence*, February 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10431567

VI. Reflection

Looking back on our group’s literature survey experience, we went through an important journey of learning and teamwork. We encountered many problems during the process, but all team members helped each other solve the problems. Each of us read five articles, which helped us better understand the n-gram model in natural language processing. We also read the reasons for others’ writing to ensure that all articles were covered under a main theme. Reviewing each other’s work was a common process in our group, especially in the second half of the writing process, when we would read others’ writing and make suggestions for revisions. However, before we officially started writing, we worked out the structure of the article together, such as what the subtitle of each section should be. We refined our subtitles based on the points that could be analyzed in each article. We faced challenges such as time conflict and arguments based on different opinions, especially since some of our team members were not in the same time zone, and it was difficult to find a time for a meeting. However, we were willing to communicate and compromise, ensuring that there would be regular meetings. We also communicated through Discord, and everyone responded quickly and raised their needs in a timely manner. We overcame these challenges through good communication and mutual respect. Our tasks were well coordinated, and everyone was focused on their own work. By revising each other’s work, we made the investigation more cohesive and high-quality. This process not only improved our academic skills, but also strengthened our teamwork and problem-solving abilities.