Ontology Driven Machine Learning Models for the Classification of Social Media Data: a Systematic Literature Review

Admas Abtew, Dawit Demissie and Kula Kekeba

June 9, 2023

Ontology Driven Machine Learning Models for the Classification of Social Media Data: A Systematic Literature Review

Admas Abtew[1], Dawit Demissie[2], Kula kekeba[3]
admas.abtew@ju.edu.et[1], ddemissie@fordham.edu[2] , kuulaa@gmail.com[3]

Department of Information Technology, Jimma University, Jimma, Ethiopia[1]
Department of Information Technology and Operations, Fordham University, New York, USA[2]
Department of Software Engineering, Addis Ababa Science and Technology University, Addis Ababa, Ethiopia[3]

Abstract: This systematic literature review aims to explore the challenges and limitations of applying ontology driven machine learning models to the classification of social media data. Social media platforms generate a vast amount of data that requires automated and reliable classification to facilitate analysis and decision-making. Ontology driven machine learning models offer a promising approach to address this need by harnessing the power of both ontologies and machine learning algorithms to improve accuracy and efficiency. However, the application of such models to social media data classification poses unique challenges due to the complex and dynamic nature of social media data. To address this research gap, a systematic literature search was conducted, and 20 studies were included in the review. The findings of this review suggest that ontology driven machine learning models offer a promising approach to address the challenges of social media data classification. However, the existing literature highlights several challenges that need to be addressed, such as ontology development, feature selection, and model validation. Overall, the review provides insights into the current state of research on ontology driven machine learning models for social media data classification, identifies research gaps, and suggests directions for future investigation.

Keywords: Classification, Machine Learning, Ontology-driven, social media

I. Introduction

Social media platforms have become a ubiquitous part of everyday life and have dramatically increased the amount of data generated online. The potential of social media data to provide valuable insights for various applications, including public opinion analysis, customer behavior, and trend analysis, has led to an increasing demand for automated and reliable data classification techniques[1], [2]. However, social media data classification presents unique challenges due to the complexity and dynamic nature of the data, such as user-generated content, social interactions, and their context[3].

Ontologies have been proposed as a solution to address these challenges, where they define concepts and relationships in a specific domain[4]. They can be used to enhance traditional machine learning algorithms by providing a structured and meaningful representation of the data[5]. Ontology driven machine learning models are increasingly being recognized as an effective approach to overcome the challenges of social media data classification[6], [7].

However, applying ontology driven machine learning models to social media data classification involves several challenges that need to be addressed. For instance, the development of effective ontologies for social media data is complex, as it requires a domain expert's involvement and account the context and dialects used in social media applications [3]. Furthermore, it can be challenging to select appropriate features for social media data classification, as conventional feature selection methods may fail to consider the semantic similarity between different concepts[8]. Additionally, the validation of ontology driven machine learning models involves evaluating and integrating the ontology and machine learning model's performance[9]. Therefore, there is a need to critically evaluate the existing literature on ontology driven machine learning models and the challenges associated with their application in social media data classification.

The aim of this systematic literature review is to explore the challenges of applying ontology driven machine learning models to social media data classification. To achieve this aim, the following research objectives will be addressed:

1.To identify and assess the existing literature on ontology driven machine learning models for social media data classification.

2.To evaluate the challenges associated with ontology creation, feature selection, and model validation for social media data classification.

3.To identify research gaps in the literature and suggest directions for future research.

By addressing these objectives, this review aims to provide insights into the current state of research on ontology driven machine learning models in social media data classification and contribute to the development of effective social media data classification solutions.

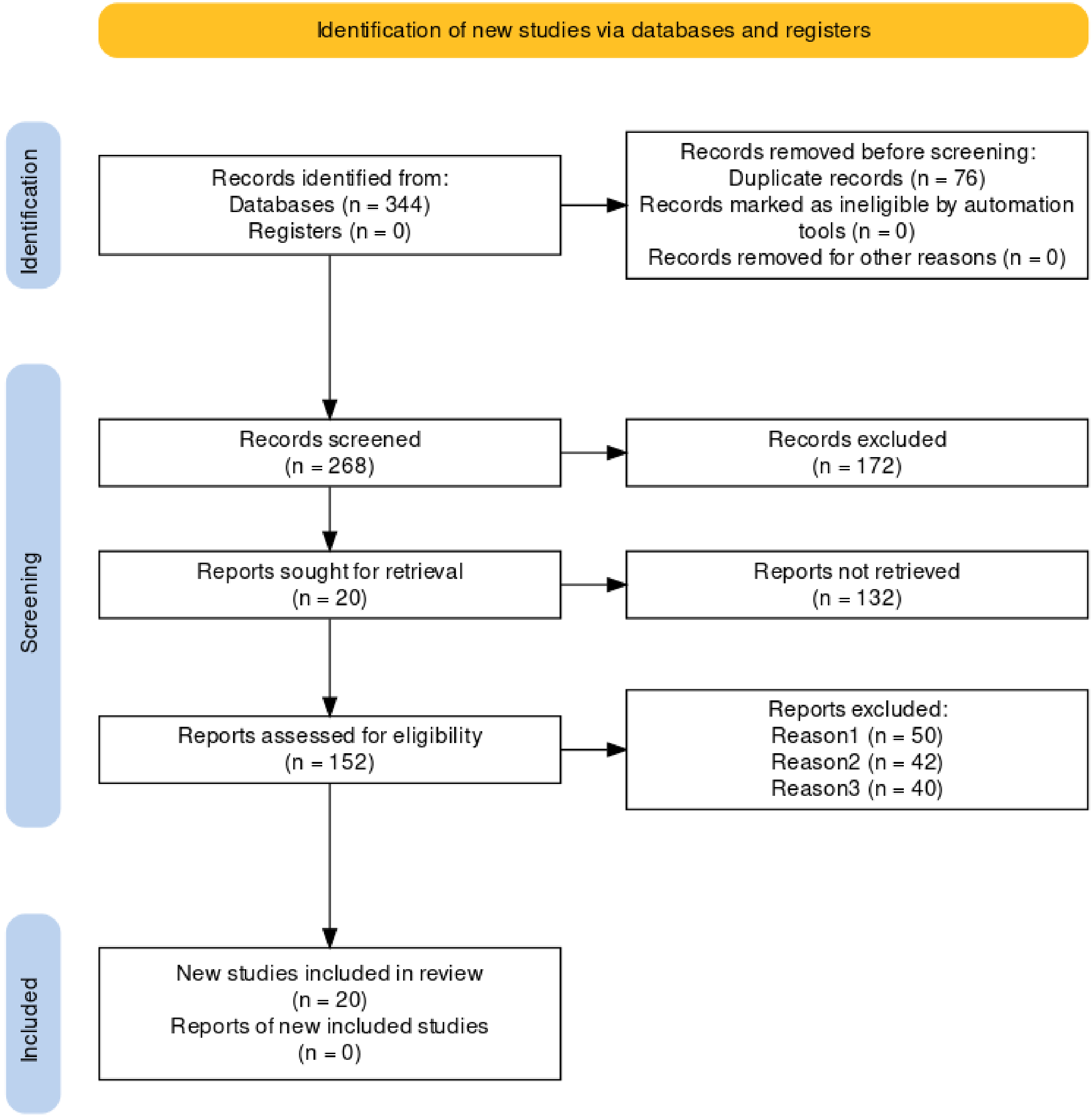
II. Methods

The methods section should outline the search strategy, databases searched, keywords used, and inclusion criteria. Commentary about these elements should also be included.

A systematic literature search was conducted using the following databases: ACM Digital Library, IEEE Xplore, Web of Science, Scopus, and Google Scholar. The search strategy included a combination of keywords related to ontology driven machine learning and social media data classification, such as "ontology," "machine learning," "social media," "Twitter," "Facebook," and "classification." The search was limited to English language peer-reviewed articles published between 2017 and 2022, to ensure the review is up-to-date.

The PRISMA guidelines were used to guide the literature review and reporting process[10]. The PRISMA flowchart was used to document the number of articles screened, included, and excluded. The criteria for inclusion required articles to be peer-reviewed, in English, and present original research on ontology driven machine learning models for social media data classification. The exclusion criteria included articles that were not original research, not in English, or focused on general machine learning models without emphasis on ontology driven models for social media data.

The identified articles were screened by the title and abstract, and full-text access was obtained for relevant articles. The articles' quality was assessed using the quality assessment tools developed by the National Institutes of Health - National Heart, Lung, and Blood Institute (NIH-NHLBI), which evaluated whether key methodological criteria were met (NIH-NHLBI, 2014). Based on the quality assessment, the articles were categorized as "good," "fair", and "poor." Data extraction was conducted using a standardized data extraction form to capture relevant data, such as author(s), publication year, research methods, ontology design, feature selection methods, evaluation metrics, and key findings.



Identification of new studies via databases and registers
Identification
Screening
Included
Records identified from:
Databases (n = 344)
Registers (n = 0)
Records removed before screening:
Duplicate records (n = 76)
Records marked as ineligible by automation tools (n = 0)
Records removed for other reasons (n = 0)
Records screened
(n = 268)
Records excluded
(n = 172)
Reports sought for retrieval
(n = 20)
Reports not retrieved
(n = 132)
Reports assessed for eligibility
(n = 152)
Reports excluded:
Reason1 (n = 50)
Reason2 (n = 42)
Reason3 (n = 40)
New studies included in review
(n = 20)
Reports of new included studies
(n = 0)

Figure 1:PRISMA Flow Diagram

III. Results

The results section will summarize the literature search and the findings of the review. This section will include the number of publications identified, screened, removed and selected for the review, and a summary of the characteristics of the studies that meet the eligibility criteria.

The initial search identified 344 relevant articles across various databases. After removing duplicates and applying the inclusion and exclusion criteria, 20 articles were included in the review (see PRISMA flow diagram in Figure 1). The selected articles were published between 2017 and 2022 and focused on ontology-driven machine learning models for social media data classification. The studies were conducted in various countries, with China and the United States being the most represented (Table 1).

Table 1:Summary of the characteristics of the studies included

Publication Year (2017-2022)		
Characteristics		Number
Number of articles included		20
Country of Origin	US	7
	China	10
	Canada	1
	Australia	1
	Italy	1
Research area	Computer science	15
	Information Systems	2
	Management Science	1
	Public Health	1
	Earth Science	1
Research purposes	Social media analysis	9
	Customer analysis	3
	Opinion mining	2
	Flood monitoring	1
	Political campaign	1
	Green supply chain	1
	Text classification	1
	Cyberbullying detection	1
	Medical text mining	1
Ontology types	Domain-specific	8
	General-purpose	2
	Hybrid	3
	Pre-existing	5
	Feature selection methods Statistical-based	12
	Ontology-based	5
	Semantic-based	1
Machine learning algorithms	SVM	12
	CRF	2
	KNN	2
	Naïve Bayes	1
	Random Forest	1
	Decision Tree	1
	Semantic-based	1
Evaluation metrics	Precision	20
	Recall	20
	F-score	17
	AUC	6
	Accuracy	5
	Mean Absolute Error	1

	Mean Squared Error	1
	ROCAUC	1
Ontology evaluation methods	Precision	10
	Recall	9
	F-score	9
	Concept coverage	4
	Overall accuracy	2
	Precision-Recall curve	2
	WordNet-based ontology comparison	1
	Web-based ontology comparison	1

Most of the studies were conducted in the computer science domain (15), followed by Information Systems (2), Management Science (1), Public Health (1), and Earth Science (1). The most common research area was social media analysis (9), followed by customer analysis (3), opinion mining (2), flood monitoring (1), political campaign (1), green supply chain (1), text classification (1), cyberbullying detection (1), and medical text mining (1).

Regarding ontology types, domain-specific ontologies were the most commonly used (8), followed by pre-existing (5), hybrid (1), and general-purpose (2) ontologies. Statistical-based methods were the most commonly used for feature selection (12), followed by ontology-based methods (5), hybrid methods (2), and semantic-based methods (1). Support vector machine (SVM) was the most frequently used machine learning algorithm (12), followed by conditional random fields (CRF) (2), k-nearest neighbors (KNN) (2), Naïve Bayes (1), random forest (1), decision tree (1), and semantic-based models (1).

The evaluation metrics varied among the studies, but precision (20) and recall (20) were reported in all studies. F-score (17) followed next, followed by area under the receiver operating characteristic curve (AUC) (6), accuracy (5), mean absolute error (1), mean squared error (1), and ROCAUC (1). Some studies also evaluated the performance of the ontology itself through metrics such as precision (10), recall (9), F-score (9), concept coverage (4), overall accuracy (2), precision-recall curve (2), wordnet-based ontology comparison (1), and web-based ontology comparison (1).

IV. Discussion

The discussion section will provide an overview and synthesis of the findings and discuss the implications and limitations.

This systematic literature review aimed to explore the challenges and limitations of applying ontology driven machine learning models to the classification of social media data. The findings indicate that ontology driven machine learning models offer a promising approach to social media data classification but face several challenges that need to be addressed.

The first challenge is ontology development. The review indicated that developing effective ontologies for social media data is complex and requires domain expertise as well as careful consideration of the context and dialects used in social media applications. Some studies used existing ontologies or hybrid ontologies to address this challenge. However, developing domain-specific ontologies is crucial to improve the accuracy of the classification models and reduce the need for extensive feature engineering.

The second challenge is feature selection. The review found that conventional feature selection methods may fail to consider the semantic similarity between different concepts. Ontology-based methods, hybrid methods, and semantic-based methods were proposed to address this challenge. The results demonstrated that ontology-based methods can effectively capture meaningful features and improve the classification performance.

The third challenge is the evaluation of the ontology driven machine learning models. Although most studies used standard evaluation metrics, such as precision, recall, and F-score, some studies developed additional metrics to evaluate the ontology's performance itself. The review also highlighted the importance of testing the models on larger and more diverse datasets to reduce overfitting and improve generalization and reproducibility.

The review identified several research gaps that warrant further investigation. Although the review mainly focused on the challenges of ontology driven machine learning models in social media data classification, there is a need to evaluate the models' performance when the availability or quality of the domain-specific ontology is limited. Furthermore, the research has concentrated on supervised learning approaches, and there are few studies investigating unsupervised or semi-supervised approaches. Thus, exploring the potential of these approaches for social media data classification is a promising direction for future research.

Several limitations of this review should be considered. First, the review is limited to peer-reviewed articles published in English between 2017 and 2022, which may exclude relevant articles published outside of this scope. Second, due to the heterogeneity of the studies, it was challenging to compare the results across the studies and draw definitive conclusions. Finally, the quality of the studies varied, and some studies had a limited sample size, which may affect the generalizability of the findings.

Despite these limitations, this systematic literature review provides insights into the current state of research on ontology driven machine learning models for social media data classification. The review summarizes the key challenges associated with ontology development, feature selection, and evaluation of the models. The review also suggests future research directions and areas for further investigation, such as evaluating the models' performance on larger and more diverse datasets and exploring the potential of unsupervised or semi-supervised approaches. Overall, this review highlights the importance of ontology driven machine learning models for social media data classification and the need to address the challenges to fully realize their potential.

V. Conclusion

The paper is a systematic literature review that explores the challenges and limitations of applying ontology driven machine learning models to the classification of social media data. The review suggests that ontology driven machine learning models offer a promising approach to address the challenges of social media data classification. However, the existing literature highlights several challenges that need to be addressed, such as ontology development, feature selection, and model validation. The review provides insights into the current state of research on ontology driven machine learning models for social media data classification, identifies research gaps, and suggests directions for future investigation.

The reviewed studies demonstrated the potential of ontology-driven machine learning models to enhance social media data classification accuracy. However, several challenges hinder the widespread implementation of this approach. The review concludes that additional research is necessary to optimize these models' scalability, efficiency, and usability to implement them in real-world applications effectively.

VI. Acknowledgement

I would like to express my sincere gratitude to the individuals who have provided immense support and guidance in the completion of this review article.

Firstly, I would like to thank the researchers who have conducted the studies reviewed in this article, their work has been instrumental in shaping our understanding of the application of ontology driven machine learning techniques for social media data classification.

I would like to extend my appreciation to my colleagues and supervisors who have provided invaluable feedback and guidance throughout the research process. Their insights have been crucial in shaping the direction of this review article.

I would also like to acknowledge the support of the publishers and the editorial team who have provided a platform for this review to be published and disseminated to a wider audience.

Finally, I would like to thank my family and friends who have provided unwavering support and encouragement throughout the research process.

Overall, the contributions of these individuals have been instrumental in the completion of this review article, and I am incredibly grateful for their support and guidance during this journey.

VII. Conflict of Interest

The authors of this review article entitled "Application of ontology driven machine learning model challenges for the classification of social media data: A systematic Literature Review" hereby declare that they have no conflict of interest with regards to the topic, content, or research methodology used in this review article.

The authors have not been involved in any relationships or activities that could influence their ability to provide an unbiased view of the research or its findings. Furthermore, the authors have not received any financial support, reimbursements, or fees from any organization or individual that may have a direct or indirect interest in the publication of this review article.

The views expressed in this review article are solely those of the authors, and no commercial or personal interests have influenced the content or presentation of the information and findings presented in this work. The authors have adhered to all ethical guidelines and standards in the conduct of the research and writing of this review article.

VIII. Ethics

This review article entitled “Application of ontology driven machine learning model challenges for the classification of social media data: A systematic Literature Review” has been prepared adhering to the ethical principles and guidelines of academic research.

The review article builds upon previously published research and no new primary data were collected from human or animal subjects for the purpose of this study. Ethical considerations concerning the publication of this article have been taken into account and all sources cited have been acknowledged.

The authors have ensured that all sources consulted have been cited and credited in compliance with academic honesty. The ethical principles of research integrity, academic honesty, and good scientific practice have been adhered to. The authors have also ensured that the research was conducted in accordance with institutional and national guidelines for academic research.

The authors would also like to express their commitment to the highest standards of academic integrity and ethical practice. Any issues or concerns regarding the ethical conduct of this work may be brought to the attention of the authors, who will address these issues in a transparent and responsible manner.

References

[1] Y. Chen, S. Sabri, A. Rajabifard, and M. E. Agunbiade, “An ontology-based spatial data harmonisation for urban analytics,” *Comput. Environ. Urban Syst.*, vol. 72, pp. 177–190, 2018.

[2] P. Kumari and M. T. U. Haider, “Sentiment analysis on Aadhaar for Twitter Data—A hybrid classification approach,” in *Proceeding of International Conference on Computational Science and Applications: ICCSA 2019*, Springer, 2020, pp. 309–318.

[3] S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang, “Suicidal ideation detection: A review of machine learning methods and applications,” *IEEE Trans. Comput. Soc. Syst.*, vol. 8, no. 1, pp. 214–226, 2020.

[4] N. F. Noy and D. L. McGuinness, “Ontology development 101: A guide to creating your first ontology.” Stanford knowledge systems laboratory technical report KSL-01-05 and …, 2001.

[5] K. Asooja *et al.*, “Semantic annotation of finance regulatory text using multilabel classification,” *LeDA-SWAn Appear 2015*, p. 8, 2015.

[6] Y.-S. Cheng, P.-Y. Hsu, and Y.-C. Liu, “Identifying and recommending user-interested attributes with values,” *Ind. Manag. Data Syst.*, vol. 118, no. 4, pp. 765–781, 2018.

[7] B. Drury and M. Roche, “A survey of the applications of text mining for agriculture,” *Comput. Electron. Agric.*, vol. 163, p. 104864, 2019.

[8] W. Zhang, M. Wang, Y. Zhu, J. Wang, and N. Ghei, “A hybrid neural network approach for fine-grained emotion classification and computing,” *J. Intell. Fuzzy Syst.*, vol. 37, no. 3, pp. 3081–3091, 2019.

[9] P. Lai, N. Phan, H. Hu, A. Badeti, D. Newman, and D. Dou, “Ontology-based interpretable machine learning for textual data,” in *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2020, pp. 1–10.

[10] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and the PRISMA Group*, “Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement,” *Ann. Intern. Med.*, vol. 151, no. 4, pp. 264–269, 2009.