ASSESSING THE EFFICACY OF MACHINE LEARNING APPROACHES IN CLASSIFYING AI RESEARCH PAPERS

Juan Camilo Rojas Lucero
Department of Computing
Boise State University
juancamilorojasl@u.boisestate.edu

Abstract

The rapid expansion of artificial intelligence (AI) research has resulted in a vast volume of academic publications, creating a pressing need for efficient methods to organize and classify these documents. This paper investigates the application of text mining techniques combined with machine learning algorithms to automate the classification of AI research papers. We specifically compare the performance of Random Forest and Support Vector Machines (SVM) to determine their effectiveness in categorizing a diverse dataset of AI literature. Our approach emphasizes the role of feature extraction and selection, particularly using term frequency-inverse document frequency (TF-IDF), to enhance the classification accuracy. Experimental results reveal that Random Forest outperforms SVM, achieving the highest accuracy due to its ability to handle high-dimensional data and complex decision boundaries. The superior performance of Random Forest underscores its suitability for classifying AI research papers, offering a robust solution for managing the ever-growing corpus of AI literature. These findings provide valuable insights into the development of automated classification systems, which are crucial for improving information retrieval and facilitating a deeper understanding of AI research trends.

1 Introduction

In the digital age, the exponential growth of academic publications in artificial intelligence (AI) has created a pressing need for efficient information organization and classification. With vast amounts of research papers emerging across various AI subfields, effective methods for categorizing and retrieving these documents are urgently required. This necessity is particularly pronounced due to the increasing complexity and diversity of AI research, where overlapping topics and methodologies complicate the landscape.

Text mining, combined with Natural Language Processing (NLP), data mining, and machine learning techniques, offers powerful tools for automating the classification of AI-related papers (Bujang et al., 2021) These technologies facilitate automatic retrieval, categorization, and summarization, enabling researchers to manage the overwhelming volume of literature. The primary goal of text classification in this context is not only to extract meaningful information from academic papers but also to ensure proper annotation and efficient representation of documents. Selecting appropriate classifier functions is crucial for maintaining generalization and avoiding overfitting, especially as the scope of AI research continues to expand (Gong et al., 2023). Text classification is a fundamental task within text mining, involving the assignment of research papers to predefined categories based on their content. This process can be single-label (assigning one category) or multi-label (assigning multiple categories), serving as a foundation for organizing AI literature across various domains(Ahmed et al., 2013). Depending on the complexity of the classification task, problems may range from binary classification to multi-class categorization, necessitating sophisticated machine learning models that learn from pre-labeled datasets. These classifiers play a pivotal role in tasks such as information retrieval, question-answering, and summarization, allowing efficient navigation through extensive digital archives of AI research.

Several machine learning algorithms have demonstrated success in classifying AI papers (Ahmed et al., 2013), including Random Forest, k-Nearest Neighbors (k-NN), and Support Vector Machines (SVM). Random Forest, an ensemble method, excels at handling high-dimensional data and mitigating overfitting, while k-NN provides a straightforward approach to categorizing documents based on similarity Sulova et al.. SVM is particularly effective in constructing optimal de-

cision boundaries for both binary and multi-class classification problemsTong and Koller, 2002. Numerous studies have highlighted the effectiveness of these algorithms in the context of AI paper classification, showcasing their ability to accurately categorize research articles and enhance literature organization.

A critical aspect of effective text classification is feature extraction, which transforms raw data into informative characteristics that capture underlying patterns. The quality and relevance of these features significantly impact algorithm performance. For instance Wan et al., 2012, SVM and k-NN depend on well-structured features for accurate classification, whereas tree-based methods like Random Forest can tolerate irrelevant features but still benefit from careful feature selection. Dimensionality reduction techniques, such as Principal Component Analysis (PCA), further enhance model efficiency by compressing the feature space while preserving essential information. In text classification, techniques like tokenization and term frequency-inverse document frequency (TF-IDF) are vital for converting textual data into numerical formats that algorithms can process effectivelyDadgar et al., 2016. Understanding this interplay is essential for optimizing machine learning models and achieving robust classification results.

Given the rapid growth and diversity of AI research, automatic text classification has become indispensable. Machine learning-based classifiers, capable of discerning distinguishing features from extensive datasets of academic papers, are essential for managing the wide range of topics within AI literature. As digital documents in this field continue to proliferate, automated classification systems offer a scalable solution for organizing research papers, facilitating improved information retrieval, and providing deeper insights into the evolving landscape of artificial intelligence.

In this paper, we present a comprehensive comparison of three prominent machine learning algorithms—Random Forest, k-Nearest Neighbors (k-NN), and Support Vector Machines (SVM)—aimed at effectively classifying the taxonomy of a database of AI papers. This analysis seeks to highlight the strengths and weaknesses of each algorithm in the context of AI literature classification, thereby contributing to the ongoing discourse on improving automated document organization.

2 Related Work

The growing abundance of textual data underscores the necessity for techniques such as text mining, machine learning, and natural language processing to effectively organize and extract patterns and insights from documents. A crucial component of this process is text representation, which has been widely explored through various statistical and syntactic approaches. The choice of representation model depends on the specific information sought, highlighting the importance of selecting the right method for the desired outcome.

The quality of the data source significantly impacts the effectiveness of classification algorithms in data mining. Irrelevant and redundant features can increase the costs of the mining process and degrade the quality of the results.

In 2015, ShrihariR and Desai, 2015 conducted a comparative analysis of classification techniques, finding that Support Vector Machines (SVM) and Naïve Bayes (NB) were the most accurate algorithms for text classification, with accuracies of 90.21

Subsequent studies have further explored the performance of these algorithms. Syahputra and Wibowo, 2023 compared SVM and Random Forest (RF) for detecting negative content online. They found that SVM significantly outperformed RF, achieving 97

Rubbens et al., 2023 examined SVM and RF for classifying Indonesian text, noting the unique challenges posed by language-specific grammar. Although SVM achieved the highest accuracy of 0.9648 with 10-fold cross-validation, it required less computational time compared to RF, which reached an accuracy of 0.9561 but took significantly longer. This study highlights SVM's efficiency and accuracy, particularly for language-specific applications.

Salles et al., 2018 addressed high-dimensional data classification challenges by implementing a modified random forest algorithm, Lazy NN RF, designed specifically for such tasks. Their integration of nearest-neighborhood projections with the random forest classifier demonstrated significant improvements in classification effectiveness.

Nadi and Moradi, 2019 enhanced random forest performance for high-dimensional data by increasing the number of trees while limiting each tree's depth, providing localized views of the problem and significantly improving accuracy.

Moldagulova and Sulaiman, 2017 used the k-Nearest Neighbors (k-NN) algorithm for document classification, emphasizing term vector space reduction. They showed that reducing the feature space by tenfold led to minimal accuracy loss while greatly reducing time costs.

Hmeidi et al., 2014 compared SVM and another machine learning model for Arabic text categorization, using tf-idf for feature weighting and CHI statistics for ranking. SVM achieved favorable results in average performance, F1-score, and prediction timing.

Hmeidi et al., 2015 developed a refinement strategy called DragPushing to address model misfitting in k-NN classifiers, significantly enhancing k-NN performance in text classification tasks.

Dadgar et al., 2016 applied the TF-IDF Vectorizer method within the k-NN framework, examining its impact on classification speed and quality, highlighting the strengths and weaknesses of the algorithm.

Gong et al., 2023 improved k-NN by combining it with a constrained one-pass clustering algorithm, significantly reducing text similarity computations and outperforming models like Naïve Bayes, SVM, and state-of-the-art k-NN classifiers.

Overall, studies such as Wan et al., 2012 emphasize the need to evaluate various classification techniques, including comparisons of feature selection methods and the performance of different algorithms. While SVM often excels, other classifiers like k-NN and Naïve Bayes also show strong potential, particularly when well-preprocessed, making it essential to consider a diverse range of methods for effective text categorization.

3 Methodology

The methodology section details the approach used to evaluate various text classification algorithms, including SVM, and Random Forest. Key steps involve data preparation through preprocessing, feature extraction with TF-IDF. The algorithms are trained and tested using metrics like accuracy, precision, and recall. Additionally, a variable selection was developed to encounter the critical variables highly related to the accuracy. Comparative experiments are conducted under consistent conditions to assess each model's performance, highlighting their strengths and limitations in text classification tasks.

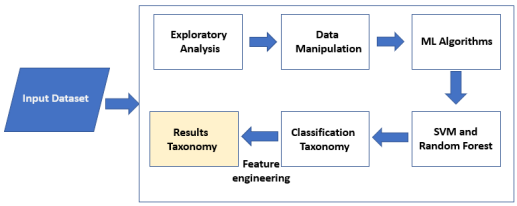

Figure 1: Methodology

3.1 Data Exploration

The dataset comprises 144 academic papers that examine advancements and trends in AI-generated content and large language models. Each entry includes essential information, such as the paper's title, authors, release date, and a direct link to the full text on arXiv. The taxonomy categorizes the papers, emphasizing their comprehensive nature and focus on topics such as generative AI, language model behavior, and large-scale language modeling. It captures key research areas including artificial intelligence, computational linguistics, and machine learning, reflecting the interdisciplinary scope of these studies. Summaries of each paper provide insights into their main findings and contributions, offering a broad perspective on the evolution, challenges, and future directions of AI-generated content and language models. This dataset serves as a critical resource for understanding current developments and guiding further research in this rapidly evolving field.

The dataset reveals a diverse distribution of research topics within AI-generated content, language models, and their associated fields. A significant portion of the papers focuses on trustworthy practices, comprehensive surveys, and prompting, underscoring the ongoing interest in exploring the capabilities and behaviors of large language models and generative AI technologies. In contrast, taxonomies related to areas such as law and finance remain underexplored, indicating a potential gap in the literature.

In recent years, there has been a noticeable increase in the number of published papers, reflecting the growing scholarly interest and rapid advancements in AI-generated content and large language models. The graph illustrates the number of papers published over specific months from July 2021 to January 2024. In July 2021, only one paper was published, indicating a slow start to research activity. This number gradually increased, with a significant rise in publications by November 2023,

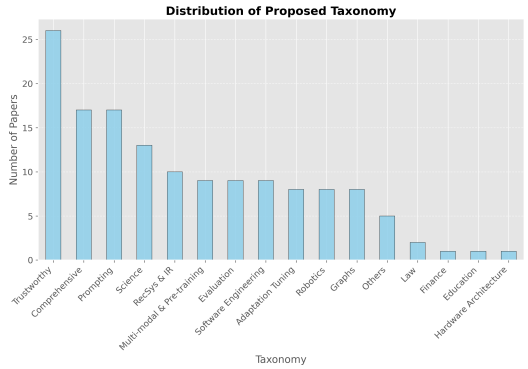

Figure 2: Distribution of taxonomies

when 21 papers were published. By January 2024, the total number of papers reached 15, suggesting a sustained interest in the topic and a growing trend in research output.

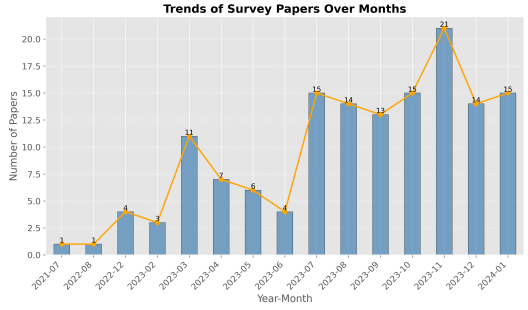

Figure 3: Trends of survey papers over months

Additionally, the graph displays the distribution of papers across various categories, highlighting the prevalence of specific fields within the research landscape. The categories Cs.CL (Computer Science - Computation and Language) and Cs.AI (Computer Science - Artificial Intelligence) show the highest values, indicating that these areas are at the forefront of current research activity. This trend suggests a strong focus on language processing and AI technologies within the academic community. Other categories, while represented, exhibit significantly lower paper counts compared to Cs.CL and Cs.AI, underscoring the dominance of these fields in the publication landscape.

On the other hand, the independence of categorical variables is crucial in statistical analysis, and the Chi-Square test serves as an effective method for assessing whether two variables are associated or if their distributions are independent, enabling researchers to uncover significant relationships within their data. The Chi-square test results highlight several relationships between various categories, showcasing significant findings. Among

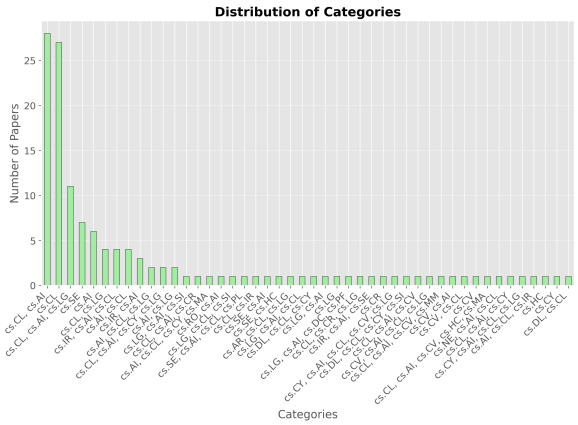

Figure 4: Distribution of categories

the comparisons, the relationship between "Comprehensive" and "Trustworthy" yielded a p-value of 0.084, indicating a notable association at a significance level of 0.1. Similarly, the connection between "Prompting" and "Trustworthy" produced the same p-value of 0.084, suggesting that trustworthiness is a crucial factor in both contexts. Furthermore, the association between "Trustworthy" and "Science" had a p-value of 0.163, which, while not below the conventional threshold, still reflects a potential link worth considering. Simultaneously, the correlation matrix heatmap aligns well with the previous chi-square results, confirming the independence of the predictors.

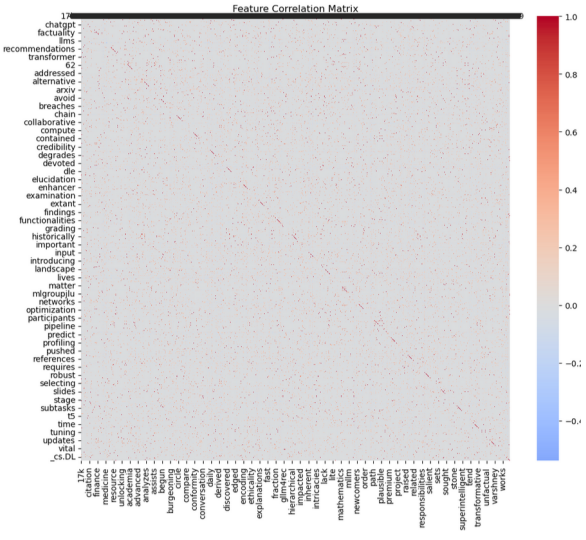

Figure 5: Correlation Matrix Heatmap

3.2 Data Manipulation

Effective data manipulation is essential for constructing a comprehensive feature matrix. In this context, missing data in critical columns, specif-

ically 'Categories' and 'ID,' were identified and addressed. An imputation technique was employed to fill these gaps, assigning the most frequent category to missing entries and generating unique identifiers for each document. This approach ensured the integrity and completeness of the dataset, which is vital for accurate analysis. Following the resolution of the missing data, the feature matrix was built by vectorizing the textual data and one-hot encoding the categorical variables. The resulting comprehensive matrix provides a robust foundation for subsequent analyses and modeling efforts.

3.3 Models Implementations

In this study, predictions were made using two widely recognized machine learning models: Random Forest (RF) and Support Vector Machine (SVM). According to the literature, both models are effective for classification tasks, particularly in handling complex datasets with high dimensionality.

Random Forest was implemented as an ensemble learning technique, leveraging multiple decision trees to improve predictive accuracy and control overfitting. The use of varying numbers of estimators (50, 100, and 200) was particularly informative, allowing for an examination of how the model's performance fluctuates with different ensemble sizes. The results indicated accuracies of 0.3276 for 50 estimators, 0.4483 for 100 estimators, and 0.4138 for 200 estimators, reflecting the impact of estimator variability on model performance.

SVM was also employed to evaluate its effectiveness in this classification context. It is well-regarded for its ability to create hyperplanes that effectively separate classes in high-dimensional spaces, with a strong capacity for handling non-linear relationships. The accuracy of the SVM model was found to be 0.3103.

4 Results and Discussion

Comparing these two models provided valuable insights into their respective strengths and weaknesses, emphasizing the importance of model selection in machine learning applications. This analysis contributes to the ongoing discourse on the effectiveness of these algorithms in predictive analytics.

The comparison of these two models provides valuable insights into their respective strengths and weaknesses, emphasizing the importance of model selection in machine learning applications.

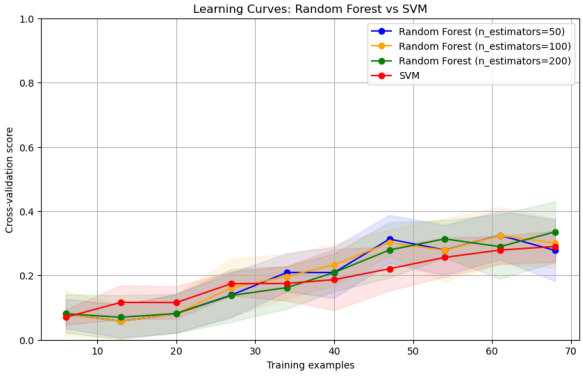

Figure 6: Machine Learning Performances

The observed lower performance of the models highlights the challenges associated with overfitting and data imbalance within the dataset. Despite efforts to enhance accuracy through model selection and parameter tuning, the results indicate that both the Random Forest and SVM models struggled to generalize effectively. Overfitting, where the models learn noise and patterns specific to the training data rather than underlying trends, can significantly compromise predictive performance. Additionally, the imbalance in class distribution may have further hindered the models' ability to classify minority classes accurately, leading to suboptimal outcomes. A potential option for improvement is to apply variable selection using the feature importance algorithm from Random Forest. This approach can help identify and retain only the most relevant features, thereby reducing model complexity and mitigating overfitting, while enhancing the models' predictive capabilities on imbalanced datasets.

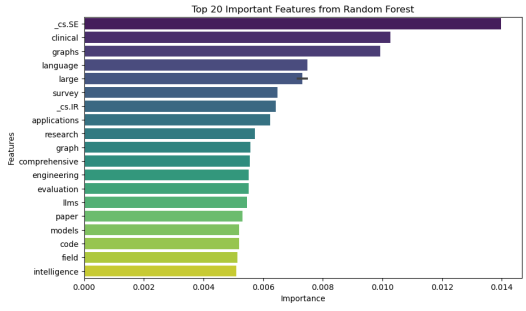

Figure 7: Variable importance

The feature importance analysis highlights the key variables that significantly contribute to predicting taxonomy. Among the top variables identified, they emerge as the most influential predictors. These features play a critical role in the model's decision-making process, indicating their relevance

in the context of the dataset. Understanding the importance of these variables can provide valuable insights into the factors that drive taxonomy classification, enhancing the overall interpretability and effectiveness of the predictive model.

Utilizing the identified key variables, the Random Forest model was retrained to evaluate its performance based on the most significant predictors. The model aims to improve its predictive accuracy for taxonomy classification. This targeted approach allows for a more refined analysis, potentially enhancing the model's ability to capture the nuances within the dataset and leading to better overall results in classifying the taxonomy effectively. The subsequent evaluation will provide insights into the model's performance with this refined feature set.

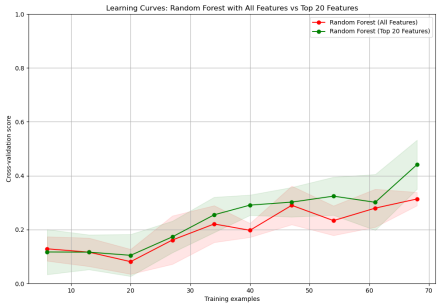

Figure 8: Variable importance

The results demonstrate a significant improvement in model performance following the selection of key variables. The accuracy of the Random Forest model when utilizing all available features was 0.3793, indicating a limited capacity to generalize from the data. However, by narrowing the focus to the top 20 most important features, the accuracy increased to 0.4828. This enhancement suggests that the selected variables provide more relevant information for the classification task, ultimately leading to better predictive performance. The findings underscore the importance of effective feature selection in improving model accuracy and achieving more reliable results in predictive analytics.

5 Conclusion

In conclusion, this paper highlights the importance of leveraging text mining techniques and machine learning algorithms to automate the classification of AI research papers, a task made increasingly essential by the rapid expansion of AI literature. Through comprehensive experimentation, we established that Random Forest outperforms Support Vector Machines in accurately categorizing a diverse dataset of AI literature, emphasizing the significance of effective feature extraction and selection, particularly utilizing the term frequency-inverse document frequency (TF-IDF) approach. The results underscore the potential of automated classification systems in enhancing information retrieval and facilitating a deeper understanding of research trends. Furthermore, our findings reveal the critical role of data exploration and manipulation in optimizing model performance. By carefully preparing and refining datasets, we can ensure that the features extracted truly represent the underlying patterns in the data, which is vital for the success of any classification task. This study provides valuable insights for future research in automated document organization, illustrating the need for ongoing development in machine learning methods to adapt to the evolving landscape of artificial intelligence research.

6 REFERENCES

References

K. Ahmed, Sultan Aljahdali, and Syed Hussain. 2013. Comparative prediction performance with support vector machine and random forest classification techniques. *International Journal of Computer Applications*, 69:12–16.

Siti Dianah Abdul Bujang, Ali Selamat, Roliana Ibrahim, Ondrej Krejcar, Enrique Herrera-Viedma, Hamido Fujita, and Nor Azura Md. Ghani. 2021. Multiclass prediction model for student grade prediction using machine learning. *IEEE Access*, 9:95608–95621.

Seyyed Mohammad Hossein Dadgar, Mohammad Shirzad Araghi, and Morteza Mastery Farahani. 2016. A novel text mining approach based on tf-idf and support vector machine for news classification. pages 112–116.

Youdi Gong, Guangzhen Liu, Yunzhi Xue, Rui Li, and Lingzhong Meng. 2023. A survey on dataset quality in machine learning. *Information and Software Technology*, 162:107268.

Ismail Hmeidi, Mahmoud Al-Ayyoub, Nawaf Abdulla, Abdalrahman Almodawar, Raddad Abooraig, and Nizar A. Ahmed. 2014. Automatic arabic text categorization: A comprehensive comparative study. *Journal of Information Science*, 41:114–124.

Ismail Hmeidi, Mahmoud Al-Ayyoub, Nawaf A. Abdulla, Abdalrahman A. Almodawar, Raddad Abooraig, and Nizar A. Mahyoub. 2015. Automatic arabic text categorization: A comprehensive comparative study. *Journal of Information Science*, 41(1):114–124.

Aiman Moldagulova and Rosnafisah Bte. Sulaiman. 2017. Using knn algorithm for classification of textual documents. pages 665–671.

Abolfazl Nadi and Hadi Moradi. 2019. Increasing the views and reducing the depth in random forest. *Expert Systems with Applications*, 138:112801.

Peter Rubbens, Stephanie Brodie, Tristan Cordier, Diogo Destro Barcellos, Paul Devos, Jose A Fernandes-Salvador, Jennifer I Fincham, Alessandra Gomes, Nils Olav Handegard, Kerry Howell, Cédric Jamet, Kyrre Heldal Kartveit, Hassan Moustahfid, Clea Parcerisas, Dimitris Politikos, Raphaëlle Sauzède, Maria Sokolova, Laura Uusitalo, Laure Van den Bulcke, Aloysius T M van Helmond, Jordan T Watson, Heather Welch, Oscar Beltran-Perez, Samuel Chaffron, David S Greenberg, Bernhard Kühn, Rainer Kiko, Madiop Lo, Rubens M Lopes, Klas Ove Möller, William Michaels, Ahmet Pala, Jean-Baptiste Romagnan, Pia Schuchert, Vahid Seydi, Sebastian Villasante, Ketil Malde, and Jean-Olivier Irisson. 2023. Machine learning in marine ecology: an overview of techniques and applications. *ICES Journal of Marine Science*, 80(7):1829–1853.

Thiago Salles, Marcos Gonçalves, Victor Rodrigues, and Leonardo Rocha. 2018. Improving random forests by neighborhood projection for effective text classification. *Information Systems*, 77:1–21.

Chauhan ShrihariR and Amish Desai. 2015. A review on knowledge discovery using text classification techniques in text mining. *International Journal of Computer Applications*, 111:12–15.

Snezhana Sulova, Latinka Todoranova, Bonimir Penchev, and Radka Nacheva. Using text mining to classify research papers.

Hermawan Syahputra and Aldiva Wibowo. 2023. Comparison of support vector machine (svm) and random forest algorithm for detection of negative content on websites. 9:165–173.

Simon Tong and Daphne Koller. 2002. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.*, 2:45–66.

Chin Heng Wan, Lam Hong Lee, Rajprasad Rajkumar, and Dino Isa. 2012. A hybrid text classification approach with low dependency on parameter by integrating k-nearest neighbor and support vector machine. *Expert Systems with Applications*, 39(15):11880–11888.