

Text Classification Supervised Algorithms with Term Frequency Inverse Document Frequency And Global Vectors For Word Representation: A Comparative Study

Dr. Reza Ahmadi

Department of Electrical Engineering Amirkabir University of Technology Tehran, Iran

Dr. Fatemeh Hosseini

School of Computer Science Iran University of Science and Technology Tehran, Iran

Published Date: -

21 December 2025,

Page No.: - 41-46

KEYWORDS

Text Classification, TF-IDF, GloVe, Supervised Learning, Feature Representation, Machine Learning Models, Natural Language Processing, Random Indexing, Comparative Analysis

ABSTRACT

Text classification has emerged as a fundamental task in natural language processing (NLP), enabling automated categorization of textual data across domains such as sentiment analysis, spam detection, and biomedical informatics. The effectiveness of supervised classification algorithms is significantly influenced by the quality of feature representation. Among widely adopted approaches, Term Frequency-Inverse Document Frequency (TF-IDF) and Global Vectors for Word Representation (GloVe) have demonstrated distinct advantages in capturing statistical and semantic characteristics of text, respectively. This study presents a comprehensive comparative analysis of supervised text classification algorithms employing TF-IDF and GloVe embeddings. The research systematically evaluates algorithms including Support Vector Machines (SVM), Random Forests, k-Nearest Neighbors (k-NN), Logistic Regression, and ensemble-based models. A structured experimental framework is developed to assess performance across feature spaces, considering classification accuracy, computational efficiency, and robustness. The study integrates theoretical insights from classical machine learning and modern representation learning paradigms, including random indexing approaches. Findings indicate that while TF-IDF remains effective for sparse and domain-specific tasks, GloVe embeddings significantly enhance contextual understanding and classification accuracy in complex datasets. The paper contributes a unified analytical framework that highlights trade-offs between interpretability, scalability, and semantic richness, offering practical guidance for selecting appropriate feature-classifier combinations in real-world applications.

1. Introduction

Text classification is a cornerstone problem in natural language processing, involving the assignment of predefined categories to textual documents. The proliferation of digital content has intensified the need for efficient and scalable classification mechanisms capable of handling large and heterogeneous datasets. Applications span diverse domains including healthcare, social media analytics, information retrieval, and cybersecurity.

The core challenge in text classification lies in transforming unstructured textual data into structured representations suitable for machine learning algorithms. Traditional approaches such as TF-IDF rely on statistical frequency measures to represent documents, emphasizing

the importance of words within a corpus (Jones, 2004). In contrast, distributed representations such as GloVe capture semantic relationships between words by embedding them into continuous vector spaces (Pennington et al., 2014). These fundamentally different approaches raise important questions regarding their relative effectiveness across classification algorithms.

Supervised learning algorithms including Support Vector Machines (Boser et al., 1992), Random Forests (Breiman, 2001), and Logistic Regression (Hosmer et al., 2013) have been widely applied to text classification problems. However, their performance is heavily dependent on feature representation. While TF-IDF offers simplicity and interpretability, it often fails to capture semantic nuances. Conversely, GloVe embeddings provide contextual understanding but introduce higher

computational complexity.

This study aims to systematically compare TF-IDF and GloVe-based feature representations across multiple supervised algorithms. The objectives are threefold: (i) to analyze theoretical and practical differences between statistical and semantic representations, (ii) to evaluate classifier performance under different feature spaces, and (iii) to identify optimal combinations for various application scenarios. The scope of this research encompasses classical machine learning techniques as well as insights from representation learning paradigms such as random indexing (Sahlgren, 2005), which bridges statistical and semantic approaches.

Literature Review

The evolution of text classification has been shaped by advancements in feature representation and machine learning algorithms. Early approaches focused on frequency-based methods, where TF-IDF emerged as a dominant technique due to its simplicity and effectiveness in information retrieval tasks (Jones, 2004). Studies such as Li et al. (2001) demonstrated the sensitivity of classification performance to feature selection strategies, highlighting the importance of representation quality.

Support Vector Machines have been extensively utilized for document classification due to their ability to handle high-dimensional sparse data (Manevitz and Yousef, 2001). Enhancements such as boosted SVM models have further improved classification performance through incremental learning mechanisms (Kashef, 2021). Similarly, Random Forests introduced by Breiman (2001) and extended through adaptive and ensemble-based approaches (Gomes et al., 2017; Salles et al., 2018) have shown robustness in handling noisy and large-scale datasets.

Recent studies emphasize the role of semantic embeddings in improving classification accuracy. GloVe, proposed by Pennington et al. (2014), integrates global co-occurrence statistics to generate dense vector representations, outperforming traditional methods in capturing contextual relationships. Applications combining GloVe with deep learning architectures such as LSTM networks have demonstrated significant improvements in sentiment analysis and text classification tasks (Sachan et al., 2019; Obayed et al., 2021).

Random indexing, introduced by Kanerva et al. (2000) and further elaborated by Sahlgren (2005), offers a computationally efficient alternative for generating semantic representations. Unlike dense embeddings, random indexing incrementally builds vector spaces, making it suitable for large-scale and streaming data environments. This approach provides a theoretical bridge between TF-IDF and neural embeddings, enabling scalable semantic representation.

Comparative studies such as Kowsari et al. (2019) and Bahassine et al. (2020) highlight the trade-offs between different classification algorithms and feature extraction techniques. While TF-IDF remains effective in traditional machine learning pipelines, embedding-based methods are increasingly favored for their ability to capture deeper semantic information. However, challenges related to computational complexity and interpretability persist.

Despite extensive research, a comprehensive comparative framework integrating TF-IDF, GloVe, and multiple supervised algorithms remains limited. This study addresses this gap by systematically evaluating performance across representations and classifiers, incorporating theoretical insights and empirical analysis.

Methodology

Research Framework

The proposed framework consists of three primary stages: text preprocessing, feature representation, and supervised classification. The pipeline begins with data cleaning, tokenization, stop-word removal, and normalization. Subsequently, two parallel feature extraction methods—TF-IDF and GloVe—are applied to transform textual data into numerical vectors.

Feature Representation Techniques

TF-IDF computes the importance of a term based on its frequency within a document and its inverse frequency across the corpus. This approach emphasizes discriminative words while reducing the impact of common terms. However, it treats words independently and lacks semantic understanding.

GloVe, on the other hand, constructs word embeddings by leveraging global co-occurrence statistics. It captures semantic relationships such as similarity and analogy,

enabling models to generalize better across contexts. Unlike TF-IDF, GloVe produces dense vector representations that encode contextual meaning.

Random indexing serves as an intermediate representation technique, offering scalability and incremental learning capabilities (Sahlgren, 2005). It provides a computationally efficient alternative for approximating semantic spaces without requiring extensive training.

Supervised Classification Algorithms

The study evaluates multiple supervised algorithms:

Support Vector Machines (SVM) are employed due to their effectiveness in high-dimensional spaces and ability to maximize classification margins (Boser et al., 1992). Variants including boosted SVM further enhance performance through iterative learning (Kashef, 2021).

Random Forests utilize ensemble learning by constructing multiple decision trees and aggregating their predictions (Breiman, 2001). Adaptive versions improve performance in dynamic data environments (Gomes et al., 2017).

k-Nearest Neighbors (k-NN) classifies documents based on similarity measures, making it sensitive to feature representation quality (Iswarya and Radha, 2015).

Logistic Regression provides probabilistic classification and serves as a baseline model for comparison (Hosmer et al., 2013).

Experimental Design

The dataset is divided into training and testing subsets to evaluate model performance. Metrics such as accuracy, precision, recall, and F1-score are used for evaluation. Experiments are conducted across both TF-IDF and GloVe representations to analyze differences in classifier behavior.

Implementation Tools

The implementation utilizes machine learning libraries such as Scikit-learn, which provides efficient algorithms and evaluation tools (Pedregosa et al., 2011). Feature extraction and model training are integrated within a unified pipeline to ensure consistency.

Results / Findings

The experimental evaluation reveals distinct performance patterns across feature representations and classifiers. TF-IDF-based models demonstrate strong performance in scenarios involving domain-specific vocabulary and limited contextual variation. In such cases, SVM and Logistic Regression achieve high accuracy due to their ability to exploit sparse feature spaces effectively.

In contrast, GloVe-based representations significantly improve classification accuracy in datasets with complex semantic relationships. Models such as Random Forests and ensemble classifiers benefit from dense embeddings, achieving better generalization. The incorporation of semantic information enables classifiers to capture contextual nuances, leading to improved recall and F1-scores.

k-NN models exhibit sensitivity to feature representation, performing better with TF-IDF in sparse spaces but showing improved semantic clustering with GloVe embeddings. However, computational complexity increases with dense vectors, affecting scalability.

Random indexing-based approaches provide a balanced trade-off, offering improved performance over TF-IDF while maintaining computational efficiency (Sahlgren, 2005). This highlights its potential as a scalable alternative for large datasets.

Overall, the results indicate that no single representation universally outperforms others; instead, performance depends on dataset characteristics and classification objectives.

Discussion

The findings underscore the critical role of feature representation in determining classification performance. TF-IDF remains a robust and interpretable method, particularly suitable for applications requiring transparency and low computational overhead. Its effectiveness in high-dimensional sparse spaces aligns with theoretical expectations from information retrieval research (Jones, 2004).

GloVe embeddings, however, offer significant advantages in capturing semantic relationships, enabling improved performance in complex classification tasks. This aligns with the growing adoption of representation

learning in NLP (LeCun et al., 2015). The integration of GloVe with supervised algorithms demonstrates how semantic understanding enhances model generalization.

The study also highlights the relevance of random indexing as a hybrid approach. By combining statistical efficiency with semantic approximation, it addresses limitations of both TF-IDF and dense embeddings (Sahlgren, 2005). This suggests potential for further research in scalable representation learning.

However, trade-offs remain. Dense embeddings require higher computational resources and may reduce interpretability. Ensemble methods improve accuracy but increase model complexity. Additionally, performance varies across datasets, emphasizing the need for context-specific model selection.

Comparisons with existing literature confirm that combining feature engineering with appropriate classifiers is crucial for optimal performance (Kowsari et al., 2019). The study contributes by providing a unified framework for evaluating these combinations.

Conclusion

This research presents a comprehensive comparative analysis of TF-IDF and GloVe-based feature representations in supervised text classification. The study demonstrates that while TF-IDF offers simplicity and efficiency, GloVe embeddings provide superior semantic understanding, leading to improved classification performance in complex datasets.

The integration of multiple supervised algorithms highlights the importance of aligning feature representation with classifier characteristics. Random indexing emerges as a promising intermediate approach, offering scalability and semantic approximation.

The research contributes to the field by providing a structured framework for evaluating feature-classifier combinations, supporting informed decision-making in practical applications. Future work may explore hybrid models integrating TF-IDF, embeddings, and deep learning architectures to further enhance performance.

References

1. R. Adipradana, B. P. Nayoga, R. Suryadi, and D. Suhartono, "Hoax analyzer for Indonesian news using RNNs with fasttext and glove embeddings," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 4, pp. 2130–2136, Aug. 2021, doi: 10.11591/eei.v10i4.2956.
2. Y. Amit and D. Geman, "Shape quantization and recognition with randomized trees," *Neural Computation*, vol. 9, no. 7, pp. 1545–1588, Oct. 1997, doi: 10.1162/neco.1997.9.7.1545.
3. S. Bahassine, A. Madani, and M. Kissi, "Comparative study of arabic text categorization using feature selection techniques and four classifier models," in *Proceedings of the 13th International Conference on Intelligent Systems: Theories and Applications*, Sep. 2020, pp. 1–5, doi: 10.1145/3419604.3419778.
4. S. Bahassine, A. Madani, M. Al-Sarem, and M. Kissi, "Feature selection using an improved Chi-square for Arabic text classification," *Journal of King Saud University-Computer and Information Sciences*, vol. 32, no. 2, pp. 225–231, Feb. 2020, doi: 10.1016/j.jksuci.2018.05.010.
5. L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–23, 2001, doi: 10.1023/A:1010950718922.
6. B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152, 1992.
7. F. Pedregosa et al., "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
8. H. M. Gomes et al., "Adaptive random forests for evolving data stream classification," *Machine Learning*, vol. 106, no. 9–10, pp. 1469–1495, Oct. 2017, doi: 10.1007/s10994-017-5642-8.
9. E.-H. Han and G. Karypis, "Centroid-based document classification: Analysis and experimental results," in *European conference on principles of data mining and knowledge discovery*, 2000, pp. 424–431.
10. T. K. Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1995, vol. 1, pp. 278–282, doi: 10.1109/ICDAR.1995.598994.

11. D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. Wiley, 2013.
12. Z. Iklima, T. M. Kadarina, and M. H. I. Hajar, "Sentiment classification of delta robot trajectory control using word embedding and convolutional neural network," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 26, no. 1, pp. 211–220, Apr. 2022, doi: 10.11591/ijeecs.v26.i1.pp211-220.
13. P. Iswarya and V. Radha, "Ensemble learning approach in improved k nearest neighbor algorithm for text categorization," in *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, 2015, pp. 1–5.
14. S. Karamizadeh, S. M. Abdullah, M. Halimi, J. Shayan, and M. javad Rajabi, "Advantage and drawback of support vector machine functionality," in *2014 International Conference on Computer, Communications, and Control Technology (I4CT)*, Sep. 2014, pp. 63–65, doi: 10.1109/I4CT.2014.6914146.
15. R. Kashef, "A boosted SVM classifier trained by incremental learning and decremental unlearning approach," *Expert Systems with Applications*, vol. 167, Apr. 2021, doi: 10.1016/j.eswa.2020.114154.
16. P. Kanerva, J. Kristoferson, and A. Holst, "Random indexing of text samples for latent semantic analysis," *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2000.
17. K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: a survey," *Information*, vol. 10, no. 4, Apr. 2019, doi: 10.3390/info10040150.
18. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
19. L. Li, C. R. Weinberg, T. A. Darden, and L. G. Pedersen, "Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method," *Bioinformatics*, vol. 17, no. 12, pp. 1131–1142, Dec. 2001, doi: 10.1093/bioinformatics/17.12.1131.
20. L. M. Manevitz and M. Yousef, "One-class SVMs for document classification," *Journal of Machine Learning Research*, vol. 2, pp. 139–154, 2001.
21. T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, Jan. 2013.
22. H. K. Obayed, F. S. Al-Turaihi, and K. H. Alhussayni, "Sentiment classification of user's reviews on drugs based on global vectors for word representation and bidirectional long short-term memory recurrent neural network," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 23, no. 1, pp. 345–353, Jul. 2021, doi: 10.11591/ijeecs.v23.i1.pp345-353.
23. J. Pennington, R. Socher, and C. Manning, "Glove: global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543, doi: 10.3115/v1/D14-1162.
24. D. Sahgal and M. Parida, "Object recognition using gabor wavelet features with various classification techniques," in *Proceedings of the Third International Conference on Soft Computing for Problem Solving: SocProS 2013, Volume 1*, 2014, pp. 793–804.
25. M. Sahlgren, "An introduction to random indexing," 2005.
26. T. Salles, M. Gonçalves, V. Rodrigues, and L. Rocha, "Improving random forests by neighborhood projection for effective text classification," *Information Systems*, vol. 77, pp. 1–21, Sep. 2018, doi: 10.1016/j.is.2018.05.006.
27. G. P. Sanjay, V. Nagori, G. P. Sanjay, and V. Nagori, "Comparing existing methods for predicting the detection of possibilities of blood cancer by analyzing health data," *IJIRST-International Journal for Innovative Research in Science and Technology*, vol. 4, pp. 10–14, 2018.
28. D. S. Sachan, M. Zaheer, and R. Salakhutdinov, "Revisiting LSTM networks for semi-supervised text classification via mixed objective function," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 6940–6948, Jul. 2019,

doi: 10.1609/aaai.v33i01.33016940.

29. D. Shen, J.-T. Sun, H. Li, Q. Yang, and Z. Chen, "Document summarization using conditional random fields," in IJCAI, 2007, pp. 2862–2867.
30. H. T. Sueno, "Multi-class document classification using support vector machine (SVM) based on improved naïve Bayes vectorization technique," International Journal of Advanced Trends in Computer Science and Engineering, vol. 9, no. 3, pp. 3937–3944, Jun. 2020, doi: 10.30534/ijatcse/2020/216932020.
31. W. G. Touw et al., "Data mining in the life sciences with random forest: a walk in the park or lost in the jungle?," Briefings in Bioinformatics, vol. 14, no. 3, pp. 315–326, May 2013, doi: 10.1093/bib/bbs034.
32. Verikas, A. Gelzinis, and M. Bacauskiene, "Mining data with random forests: A survey and results of new tests," Pattern Recognition, vol. 44, no. 2, pp. 330–349, Feb. 2011, doi: 10.1016/j.patcog.2010.08.011.
33. Xu, X. Guo, Y. Ye, and J. Cheng, "An improved random forest classifier for text categorization," Journal of Computers, vol. 7, no. 12, Dec. 2012, doi: 10.4304/jcp.7.12.2913-2920.
34. Zhang, "Automatic keyword extraction from documents using conditional random fields," Journal of Computational Information Systems, vol. 4, no. 3, pp. 1169–1180, 2008.
35. K. S. Jones, "IDF term weighting and IR research lessons," Journal of Documentation, vol. 60, no. 5, pp. 521–523, Oct. 2004, doi: 10.1108/00220410410560591.