

Evaluating the Generalizability of Semantic Term Synthesis with Chi-Square Guided Feature Selection Technique across Multiple Classifiers

Simon, A.¹, Junaidu, S.B.², Aliyu, S.³ and Yahaya, A.⁴

¹ Department of Computer Science, Kaduna State University (KASU)

^{2,3} Department of Computer Science, Ahmadu Bello University Zaria

³ Department of Statistics, Ahmadu Bello University Zaria⁴

a.simon@kasu.edu.ng¹

Abstract

Feature selection is essential in text classification because it reduces dimensionality and improves the quality of document representation. However, many traditional feature selection methods rely mainly on statistical term distributions and often overlook semantic relationships among terms. This study evaluates the effectiveness of the Semantic Term Synthesis with Chi-Square (T-Sync) framework, which enhances feature representation by synthesizing semantically related terms before feature selection. Experiments were conducted on the 20 Newsgroups and Reuters-21578 benchmark datasets using four machine learning classifiers: Naïve Bayes, Support Vector Machine, K-Nearest Neighbors, and Random Forest. Their performance was evaluated using Accuracy, F1-score, and execution time across different feature dimensions. On the 20 Newsgroups dataset, T-Sync achieved average accuracies of 0.7941 (NB), 0.7917 (SVM), 0.5799 (KNN), and 0.7809 (RF). On the Reuters dataset, the corresponding average accuracies were 0.9261 (NB), 0.9391 (SVM), 0.8150 (KNN), and 0.9318 (RF). In addition to improving classification performance over Var-CV-CHI in most experimental settings, T-Sync consistently required less execution time, demonstrating its computational efficiency. These findings show that incorporating semantic term synthesis into feature selection can improve both the effectiveness and efficiency of text classification systems.

Keywords: Text Classification, Feature Selection, Semantic Term Synthesis, Chi-Square

1. Introduction

The rapid growth of digital information has led to an enormous amount of textual data being generated every day through news articles, social media platforms, emails, online forums, and web

documents (Minaee et al., 2021). Managing and organizing this information manually has become increasingly difficult, making automatic text classification an essential tool in modern information systems. Text classification enables documents to be assigned to predefined categories and is widely used in applications such as sentiment analysis, spam filtering, topic categorization, information retrieval, and recommendation systems (Aggarwal & Zhai, 2012; Kowsari et al., 2019; Minaee et al., 2021; Sebastiani, 2002). The success of any text classification system depends greatly on how textual information is represented and transformed into features that can be processed by machine learning algorithms. Traditional text representation methods often generate very large feature spaces containing thousands of terms, many of which may be irrelevant or redundant. Such high-dimensional representations not only increase computational burdens but can also reduce classification effectiveness. For this reason, feature selection has become an important stage in text classification, helping to identify the most informative features while removing unnecessary redundancies (Ashiru et al., 2025b; Forman, 2003; Liu & Motoda, 2007; Otunba et al., 2022; Yang & Pederson, 1997). Among the many feature selection techniques proposed in literature, the Chi-Square (χ^2) statistic remains one of the most widely used because of its ability to measure the relationship between terms and class labels (Cai et al., 2021; Chuanxin et al., 2015; Otunba et al., 2022; Sun et al., 2017; Y. Yang & Liu, 1999). By selecting features that are strongly associated with specific classes, Chi-Square helps reduce dimensionality while retaining important classification information (Yang & Pederson, 1997). Despite its effectiveness, conventional Chi-Square-based approaches focus primarily on individual terms and often overlook semantic relationships among related words. Consequently, the contextual meaning shared by related terms may not be fully captured, limiting the quality of the resulting document representation (Ashiru et al., 2025b; Uysal & Gunal, 2014). Chi-Square Guided Feature Selection (T-Sync) framework. The framework enhances document representation by combining semantically related terms before applying Chi-Square feature selection. This approach seeks to enrich the feature space, reduce redundancy, and improve the overall quality of selected features. Experiments reported in Ashiru et al. (2025a) showed promising improvements in classification performance when the framework was evaluated using the Support Vector Machine (SVM) classifier. Although these findings demonstrate the potential of T-Sync, the framework was evaluated using only a single classifier. Since machine learning algorithms differ in their learning strategies, assumptions, and decision-making processes, a feature engineering technique that performs well with one classifier may not necessarily produce similar results with others (Sebastiani, 2002). Consequently, the generalizability of the T-Sync framework across different classification algorithms remains an open

question. This study addresses that gap by evaluating the performance of T-Sync using four widely adopted machine learning classifiers: Support Vector Machine (SVM), Naïve Bayes (NB), K-Nearest Neighbors (KNN), and Random Forest (RF). The evaluation is conducted on two benchmark text datasets, namely the 20 Newsgroups and Reuters collections. Performance is assessed using Accuracy, Precision, Recall, and F1-score, while computational efficiency is examined through execution time analysis. By comparing T-Sync with the Var-CV-CHI feature selection approach across multiple classifiers and feature dimensions, this study provides a broader understanding of the robustness, effectiveness, and practical applicability of semantic feature synthesis for text classification.

2. Related Works

Text datasets are typically characterized by very high dimensionality, often containing thousands of unique terms. Many of these terms may be irrelevant or redundant, leading to increased computational cost and reduced classification performance (Forman, 2003). Feature selection techniques are therefore widely employed to identify the most informative features while reducing the size of the feature space, improving classification accuracy, and lowering computational complexity (Chandrashekar & Sahin, 2014; Liu & Motoda, 2007). Several feature selection methods have been proposed, including Chi-Square (χ^2), Information Gain (IG), Mutual Information (MI), Document Frequency (DF), and variance-based approaches (Ashiru et al., 2025a; Chandrashekar & Sahin, 2014; Deng et al., 2019; Otunba et al., 2022; Yang & Pederson, 1997). Among these methods, Chi-Square has remained one of the most widely adopted because of its ability to measure the statistical dependence between terms and class labels. Previous studies have shown that Chi-Square-based feature selection can effectively reduce dimensionality while maintaining or improving classification performance across different text classification tasks (Forman, 2003; Uysal & Gunal, 2014; Yang & Pederson, 1997). Although traditional feature selection techniques have achieved considerable success, they generally evaluate terms independently and often ignore semantic relationships among words. Consequently, semantically related terms may be treated as unrelated features despite conveying similar meanings (Ashiru et al., 2025b; Mikolov et al., 2013; Pennington et al., 2014). To address this limitation, researchers have explored semantic feature engineering techniques that incorporate contextual and semantic information into document representation. Examples include n-grams, phrase extraction methods, latent semantic analysis, word embeddings, and semantic feature synthesis techniques (Deerwester et al., 1990; Jurafsky & Martin, 2023; Mikolov et al., 2013). One recent development in this area

is the Semantic Term Synthesis with Chi-Square Guided Feature Selection (T-Sync) framework proposed by (Ashiru et al., 2025a). The framework synthesizes semantically related terms before applying Chi-Square feature selection, thereby generating a more expressive and informative feature space. Experimental results demonstrated improved classification performance when evaluated using the Support Vector Machine classifier. Numerous studies have compared the performance of machine learning classifiers for text classification tasks, with Support Vector Machine consistently emerging as one of the strongest performers for high-dimensional textual data (Joachims, 1998; Sebastiani, 2002). Support Vector Machine has frequently been reported as one of the most effective classifiers for high-dimensional text data because of its ability to construct optimal decision boundaries even when the number of features greatly exceeds the number of training instances (Joachims, 1998). Naïve Bayes remains widely used due to its simplicity, efficiency, and strong performance on sparse textual representations. K-Nearest Neighbors has also been successfully applied to text categorization, although its effectiveness is often influenced by the quality of the feature space. Random Forest provides robustness through ensemble learning and has demonstrated competitive performance across a variety of classification problems. While existing comparative studies have evaluated these classifiers under different feature selection methods and document representations, relatively few studies have examined how semantic feature engineering frameworks perform across multiple classification algorithms. In particular, the generalizability of the T-Sync framework beyond Support Vector Machine has not been extensively investigated. This study therefore evaluates T-Sync using Naïve Bayes, Support Vector Machine, K-Nearest Neighbors, and Random Forest classifiers to determine whether the benefits of semantic term synthesis can be consistently realized across different learning algorithms.

3. Methodology

This study adopts an experimental research design to evaluate the generalizability of the Semantic Term Synthesis with Chi-Square Guided Feature Selection (T-Sync) framework proposed by in (Ashiru et al., 2025a) across multiple machine learning classifiers. The T-Sync framework is implemented without modification to maintain consistency with the original study and ensure a fair comparison evaluation of results. While the original earlier work evaluated the framework using only the Support Vector Machine (SVM) classifier, this study extends the evaluation to three additional classifiers: Naïve Bayes (NB), K-Nearest Neighbors (KNN), and Random Forest (RF). This broader evaluation enables the assessment of whether the performance improvements

associated with T-Sync are specific to SVM or can be generalized across different classification algorithms. The experimental evaluation is conducted using two widely used benchmark text classification datasets, namely the 20 Newsgroups dataset (Lang, 1995) and the Reuters-21578 dataset (Lewis, 1997). Classification performance is measured using Accuracy and F1-score, while computational efficiency is assessed through execution time analysis. The performance of T-Sync is compared with that of the Var-CV-CHI feature selection technique across multiple feature dimensions ranging from 500 to 4,000 features. The Var-CV-CHI conceptual flow shown in figure 1 follows a systematic workflow designed to enhance traditional Chi square (χ^2) feature selection by incorporating variance-based adjustments (Cai et al., 2021). The process begins with the input corpus, which undergoes standard preprocessing steps such as tokenization, stop-word removal, and normalization to produce a refined set of feature subset. For each term, the traditional χ^2 statistic is first computed to measure its association with the target classes. The model then calculates the variance and coefficient of variation (CV) of term frequencies across all classes, capturing how consistently or unevenly a feature appears within the dataset. Using these dispersion measures, the χ^2 scores are adjusted through the parameters β , γ , ξ , and ε , which collectively account for frequency distribution, inter-class variability, and the relative importance of feature dispersion. This adjustment refines each feature's discriminative weight beyond standard χ^2 scoring. The features are then ranked according to their adjusted scores, and the top-K features are selected to construct a more stable and informative feature set. Finally, these selected features are used to train classification models, and the system's effectiveness is evaluated using standard performance metrics.

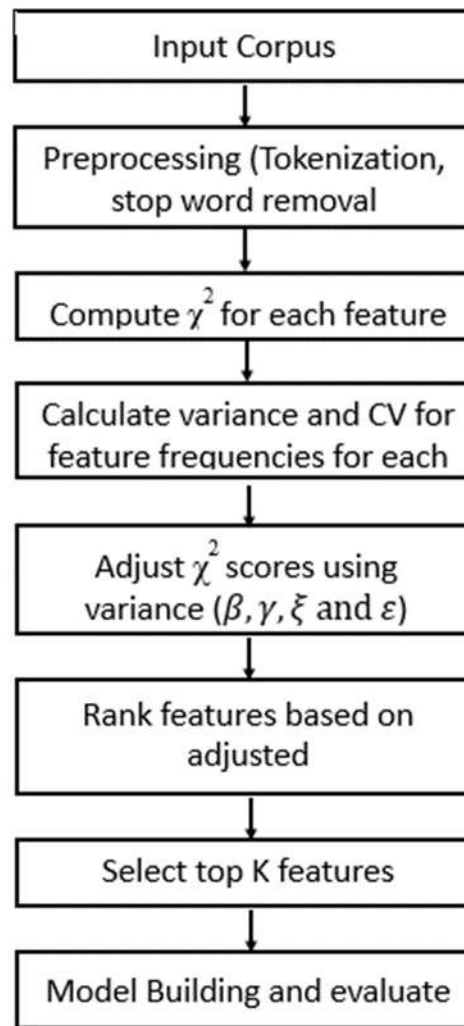


Figure 1: Conceptual Framework for Var-CV-CHI (Cai et al., 2021)

The T-Sync framework shown in figure 2 begins with the input corpus, followed by standard preprocessing steps including tokenization and stop-word removal. Term synthesis is then applied to generate meaningful features that capture semantic relationships within the text. For each feature, a χ^2 score is computed to measure its association with the target categories. Unique features (UF) are subsequently identified, and the χ^2 scores are modified based on these features to better highlight discriminative terms. The modified scores are ranked, and the top K features are selected for model building. Finally, the selected features are used to train and evaluate classifiers such as NB, SVM, KNN and Random forest.

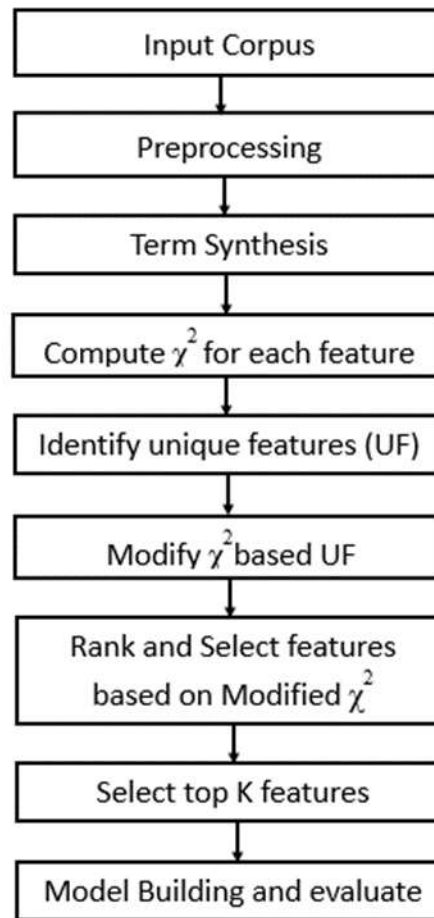


Figure 2: Conceptual Framework for T-Sync (Ashiru et al., 2025a)

To investigate the impact of feature dimensionality, experiments were conducted using feature sizes of: 500, 1000, 1500, 2000, 2500, 3000, 3500 and 4000. This range was selected to balance representational adequacy and computational efficiency while allowing systematic observation of performance trends. To evaluate the generalizability of the T-Sync framework, four widely used machine learning classifiers were employed: Support Vector Machine (SVM), Naïve Bayes (NB), K-Nearest Neighbors (KNN), and Random Forest (RF). SVM is a supervised learning algorithm that constructs an optimal separating hyperplane by maximizing the margin between classes, making it particularly effective for high-dimensional text classification tasks. Naïve Bayes is a probabilistic classifier based on Bayes' theorem that assumes conditional independence among features and is widely adopted in text classification due to its simplicity, efficiency, and strong performance on sparse textual data. K-Nearest Neighbors is an instance-based learning algorithm that classifies a document according to the majority class of its nearest neighbors in the feature space, thereby relying on similarity-based decision making. Random Forest is an ensemble learning

technique that builds multiple decision trees and combines their predictions through a voting mechanism, resulting in improved classification accuracy, robustness, and resistance to overfitting. In this study, performance evaluation metrics such as Accuracy, F-measure (harmonic means) and execution time are used to assess the proposed models.

4. Results and discussion

Table 1. Accuracy evaluation on 20newsgroup dataset

Classifier	Method	500	1000	1500	2000	2500	3000	3500	4000	AVG
NB	Var-CV-CHI	0.7453	0.7668	0.7780	0.7898	0.7903	0.7923	0.7908	0.7949	0.7810
	T-Sync	0.7841	0.7862	0.7959	0.8000	0.8010	0.7985	0.7934	0.7934	0.7941
SVM	Var-CV-CHI	0.7504	0.7637	0.7811	0.7811	0.7918	0.7964	0.7944	0.7980	0.7821
	T-Sync	0.7841	0.7872	0.7944	0.7969	0.7949	0.7934	0.7898	0.7928	0.7917
KNN	Var-CV-CHI	0.4726	0.4041	0.3458	0.3176	0.3064	0.3223	0.3217	0.3054	0.3495
	T-Sync	0.6220	0.5821	0.5340	0.5294	0.5714	0.5877	0.5923	0.6199	0.5799
RF	Var-CV-CHI	0.7586	0.7637	0.7683	0.7770	0.7698	0.7754	0.7775	0.7775	0.7710
	T-Sync I	0.7760	0.7831	0.7816	0.7785	0.7790	0.7806	0.7841	0.7841	0.7809

Table 2. F1-Socre evaluation on 20newsgroup dataset

Classifier	Method	500	1000	1500	2000	2500	3000	3500	4000	AVG
NB	Var-CV-CHI	0.7456	0.7673	0.7786	0.7901	0.7905	0.7924	0.7907	0.7949	0.7813
	T-Sync	0.7849	0.7866	0.7963	0.8003	0.8014	0.7988	0.7937	0.7937	0.7945
SVM	Var-CV-CHI	0.7504	0.7635	0.7809	0.7808	0.7917	0.7961	0.7940	0.7977	0.7819
	T-Sync	0.7842	0.7872	0.7943	0.7968	0.7947	0.7933	0.7899	0.7931	0.7917
KNN	Var-CV-CHI	0.4755	0.4018	0.3255	0.2852	0.2640	0.2868	0.2898	0.2656	0.3243
	T-Sync	0.6204	0.5842	0.5322	0.5291	0.5677	0.5887	0.5953	0.6201	0.5797
RF	Var-CV-CHI	0.7584	0.7637	0.7684	0.7770	0.7697	0.7753	0.7773	0.7775	0.7709
	T-Sync	0.7761	0.7833	0.7819	0.7789	0.7792	0.7804	0.7841	0.7840	0.7810

Tables 1 and 2 present the Accuracy and F1-score results obtained on the 20 Newsgroups dataset. The results show that T-Sync consistently outperformed Var-CV-CHI across all classifiers and feature sizes. For Naïve Bayes, T-Sync achieved an average accuracy of 0.7941 compared to 0.7810 for Var-CV-CHI. Similar improvements were observed for SVM, KNN, and Random Forest. These performance trends are visually illustrated in Figures 3–6, where the T-Sync curves generally remain above those of Var-CV-CHI across the evaluated feature dimensions. Figure 5 clearly illustrates the difference in KNN performance between T-Sync and Var-CV-CHI. Across all feature dimensions, T-Sync consistently achieved higher accuracy than Var-CV-CHI. For instance, at 500 features, T-Sync recorded an accuracy of 0.6220 compared to 0.4726 for Var-CV-CHI. The gap

became even more pronounced at 4000 features, where T-Sync achieved 0.6199 while Var-CV-CHI dropped to 0.3054. This trend explains the increase in average accuracy from 0.3495 under Var-CV-CHI to 0.5799 under T-Sync. The result suggests that incorporating semantic information into the feature selection process helps KNN identify document similarities more effectively. Among the classifiers evaluated, Naïve Bayes produced the highest average Accuracy (0.7941) and F1-score (0.7945) under T-Sync. This suggests that the semantic features generated by the proposed framework are particularly well suited to probabilistic classification. As shown in Figure 3, Naïve Bayes maintained consistently strong performance across the different feature dimensions, further demonstrating the stability of the T-Sync framework.

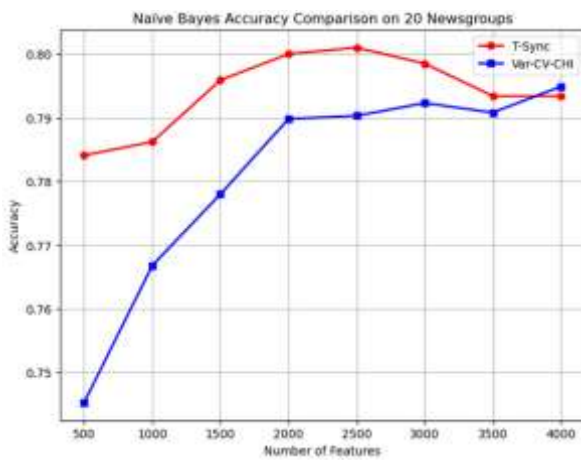


Figure 3: NB Accuracy performance

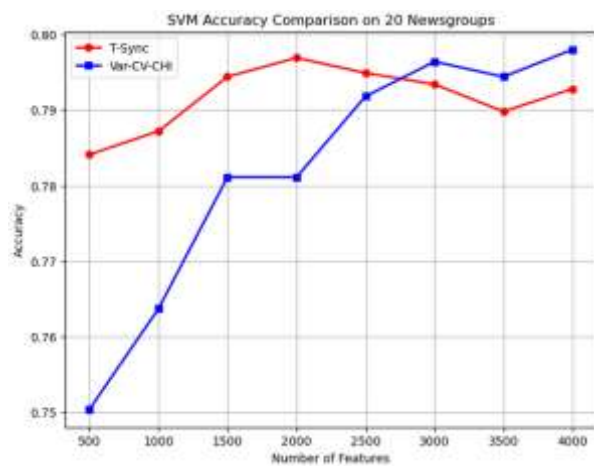


Figure 4: SVM accuracy performance

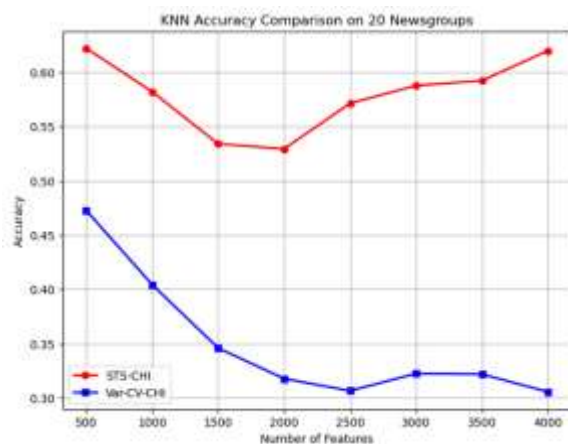


Figure 5: KNN Accuracy Performance

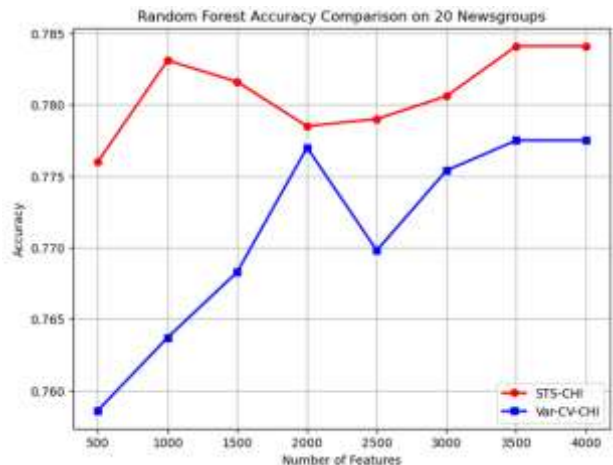


Figure 6: Random Forest Accuracy Performance

Table 3. Accuracy Comparison on Reuters Dataset

Classifier	Method	500	1000	1500	2000	2500	3000	3500	4000	AVG
NB	Var-CV-CHI	0.8572	0.8888	0.8982	0.8982	0.8915	0.8915	0.8877	0.8849	0.8873
	T-Sync	0.9179	0.9261	0.9273	0.9285	0.9289	0.9277	0.9265	0.9257	0.9261
SVM	Var-CV-CHI	0.9380	0.9430	0.9447	0.9452	0.9458	0.9480	0.9447	0.9447	0.9443
	T-Sync	0.9340	0.9387	0.9375	0.9403	0.9395	0.9387	0.9411	0.9426	0.9391
KNN	Var-CV-CHI	0.8589	0.8290	0.8152	0.8019	0.7853	0.7792	0.7576	0.7548	0.7977
	T-Sync	0.8605	0.8193	0.7917	0.8000	0.8134	0.7984	0.8114	0.8251	0.8150
RF	Var-CV-CHI	0.9286	0.9258	0.9286	0.9253	0.9253	0.9247	0.9258	0.9236	0.9260
	T-Sync	0.9324	0.9324	0.9328	0.9379	0.9293	0.9320	0.9285	0.9293	0.9318

Table 4. F1-Score Comparison on Reuters Dataset

Classifier	Method	500	1000	1500	2000	2500	3000	3500	4000	AVG
NB	Var-CV-CHI	0.8476	0.8838	0.8936	0.8926	0.8856	0.8851	0.8797	0.8765	0.8806
	T-Sync	0.9146	0.9231	0.9245	0.9255	0.9254	0.9236	0.9220	0.9210	0.9225
SVM	Var-CV-CHI	0.9371	0.9427	0.9441	0.9447	0.9453	0.9477	0.9442	0.9443	0.9438
	T-Sync	0.9325	0.9372	0.9360	0.9386	0.9377	0.9370	0.9395	0.9410	0.9374
KNN	Var-CV-CHI	0.8581	0.8267	0.8106	0.7967	0.7807	0.7744	0.7544	0.7500	0.7939
	T-Sync	0.8500	0.8054	0.7774	0.7855	0.8011	0.7866	0.8014	0.8165	0.8030
RF	Var-CV-CHI	0.9276	0.9246	0.9270	0.9237	0.9235	0.9225	0.9235	0.9212	0.9242
	T-Sync	0.9314	0.9310	0.9314	0.9366	0.9277	0.9303	0.9265	0.9272	0.9303

Tables 3 and 4 present the classification results obtained on the Reuters dataset. In contrast to the 20 Newsgroups dataset, the performance differences between T-Sync and Var-CV-CHI were not uniform across all classifiers. Overall, T-Sync achieved better results for Naïve Bayes, KNN, and Random Forest, demonstrating its ability to improve feature quality through semantic term synthesis.

For instance, Naïve Bayes recorded an average accuracy of 0.9261 with T-Sync, compared to 0.8873 obtained using Var-CV-CHI. Similarly, Random Forest achieved an average accuracy of 0.9318 under T-Sync, outperforming the 0.9260 recorded for Var-CV-CHI. KNN also benefited from the proposed approach, indicating that the synthesized semantic features helped create a more meaningful representation of document similarities. A different trend was observed for SVM. Although both methods produced strong results, Var-CV-CHI achieved a slightly higher average accuracy (0.9443) and F1-score (0.9438) than T-Sync, which recorded 0.9391 and 0.9374, respectively. This outcome suggests that SVM was already able to exploit the discriminative information contained in the original feature space effectively, leaving less room for additional gains from semantic term synthesis. Nevertheless, the overall results indicate that T-Sync remains competitive and provides consistent improvements for most classifiers on the Reuters dataset.

Table 5: Execution time table (Seconds)

Dataset	Technique	500	1000	1500	2000	2500	3000	3500	4000	AVG
20 Newsgroups	T-Sync	13.07	13.22	14.49	15.50	17.78	19.28	20.85	20.85	16.88
	Var-CV-CHI	21.82	24.15	26.41	27.98	29.92	30.49	31.34	33.06	28.15
Reuters	T-Sync	14.68	16.75	19.52	20.85	22.08	23.38	28.63	26.02	21.49
	Var-CV-CHI	21.82	32.11	37.18	36.92	38.54	37.49	40.43	42.13	35.83

Figure 7 presents the execution time comparison between T-Sync and Var-CV-CHI on the 20 Newsgroups and Reuters datasets. The results show that T-Sync consistently required less computational time than Var-CV-CHI across all feature dimensions. On the 20 Newsgroups dataset, T-Sync recorded execution times ranging from 13.07 to 20.85 seconds, whereas Var-CV-CHI required between 21.82 and 33.06 seconds. A similar trend was observed on the Reuters dataset, where T-Sync execution times ranged from 14.68 to 28.63 seconds compared to 21.82 to 42.13 seconds for Var-CV-CHI. Although the execution time of both methods increased with the number of selected features, the growth rate of T-Sync remained lower. This demonstrates that the T-Sync approach not only improves classification effectiveness but also offers better computational efficiency, making it more suitable for large-scale text classification tasks.

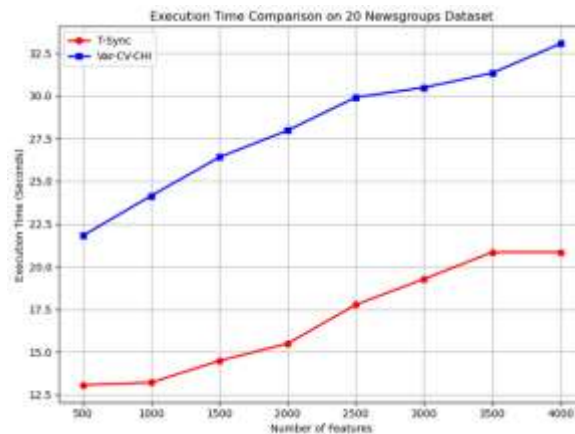


Figure 7: Execution time chart

The results demonstrate that classification performance generally improved as the number of selected features increased from 500 to approximately 2000 – 3000 features. Beyond this range, performance gains became marginal and, in some cases, slightly declined. For example, Naïve Bayes under T-Sync on the 20 Newsgroups dataset achieved its highest accuracy of 0.8010 at 2500 features, while further increases in feature size did not lead to substantial improvements. This observation indicates that excessively large feature sets may introduce redundant information without contributing meaningful

discriminative power. These findings justify the use of multiple feature sizes during experimentation and demonstrate that an optimal feature subset exists between representation quality and dimensionality reduction. The results reveal that T-Sync consistently improved performance for Naïve Bayes, KNN, and Random Forest across both datasets. Furthermore, T-Sync outperformed Var-CV-CHI for all classifiers on the 20 Newsgroups dataset. These findings indicate that the T-Sync framework is not dependent on a single classifier and can be effectively applied to different learning paradigms. Although Var-CV-CHI achieved slightly better results with SVM on the Reuters dataset, the overall results suggest that T-Sync exhibits strong classifier independence and robust generalization characteristics.

5. Conclusion

This study evaluated the effectiveness of the Semantic Term Synthesis with Chi-Square Guided Feature Selection (T-Sync) framework across multiple machine learning classifiers using the 20 Newsgroups and Reuters-21578 benchmark datasets. The results showed that T-Sync consistently improved classification performance over Var-CV-CHI, particularly on the 20 Newsgroups dataset, where it achieved higher Accuracy and F1-scores across all classifiers. The framework was especially beneficial for K-Nearest Neighbors, indicating that semantic term synthesis creates a more informative feature space that improves the ability of distance-based classifiers to identify similar documents. On the Reuters dataset, T-Sync also produced competitive results, outperforming Var-CV-CHI for Naïve Bayes, K-Nearest Neighbors, and Random Forest, while Support Vector Machine achieved slightly higher performance with Var-CV-CHI. In addition, T-Sync consistently required less execution time across both datasets, demonstrating improved computational efficiency. These findings suggest that incorporating semantic relationships into the feature selection process enhances the quality of document representation and leads to more effective text classification. The results also indicate that the benefits of T-Sync are not limited to a single classifier, demonstrating its robustness and generalizability across different machine learning algorithms. Therefore, T-Sync represents a practical and scalable feature selection framework for text classification, particularly in applications involving large and high-dimensional textual datasets.

Reference

- Aggarwal, C. C; & Zhai, C. (2012). *Mining Text Data*: Springer.
- Ashiru, S; Junaidu, S.B; Aliyu, S; & Yahaya, A. (2025a). Semantic Term Synthesis with Chi-Square Guided Feature Selection for Improved Text Classification. *Bayero Journal of Pure and Applied Sciences*, 18(2), pp. 248-255. doi: [dx.doi.org/10.4314/bajopas.v18i2.31](https://doi.org/10.4314/bajopas.v18i2.31)
- Ashiru, S; Junaidu, Sahalu; Aliyu, Salisu; & Yahaya, Abubakar. (2025b). A Review of Dimension Reduction and Feature Selection: A New Perspective. *Bima Journal of Science and Technology*, Vol. 9((2B)), pp. 371-377. doi: [10.64290/bima.v9i2B.1330](https://doi.org/10.64290/bima.v9i2B.1330)
- Cai, Liang-jing; Lv, Shu; & Shi, Kai-bo. (2021). Application of an Improved CHI Feature Selection Algorithm. *Discrete Dynamics in Nature and Society*, vol. 2021, pp. 1-8. doi: [10.1155/2021/9963382](https://doi.org/10.1155/2021/9963382)
- Chandrashekar, G; & Sahin, F. . (2014). A Survey on Feature Selection Methods. *Computers & Electrical Engineering*, 40(1), pp. 16-28. doi: doi.org/10.1016/j.compeleceng.2013.11.024
- Chuanxin, Jin; Tinghuai, Ma; Rongtao, Hou; Meili, Tang; Yuan, Tian; Abdullah, Al-Dhelaan; & Mznah, Al-Rodhaan. (2015). Chi-square Statistics Feature Selection Based on Term Frequency and Distribution for Text Categorization. *IETE Journal of Research*, 61, pp.1-12.
- Deerwester, S; Dumais, S. T; Furnas, G. W; Landauer, T. K; & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 4(16), pp. 391-407. doi: [doi.org/10.1002/\(SICI\)1097-4571\(199009\)41:6<391::AID-ASI1>3.0.CO;2-9](https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9)
- Deng, Xuelian; Li, Yuqing; Weng, Jian; & Zhang, Jilian. (2019). Feature selection for text classification: A review. *Multimedia Tools and Applications*, 78(3), pp.3797-3816. doi: [10.1007/s11042-018-6083-5](https://doi.org/10.1007/s11042-018-6083-5)
- Forman, George. (2003). An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *Journal of Machine Learning Research - JMLR*, 3, pp. 1289–1305
- Joachims, T. (1998). *Text Categorization with Support Vector Machines: Learning with many Relevant Features*. Paper presented at the Proceedings of the 10th European Conference on Machine Learning (ECML 1998)
- Jurafsky, D; & Martin, J. H. (2023). *Speech and Language Processing*: Stanford University.
- Kowsari, K; K, Jafari Meimandi; Heidarysafa, M; Mendu, S; Barnes, L. E; & Brown, D. E. (2019). Text Classification Algorithms: A Survey. *Information* 10(4). doi: doi.org/10.3390/info10040150
- Lang, K. (1995). *NewsWeeder: Learning to Filter Netnews*. Paper presented at the In Proceedings of the 12th International Conference on Machine Learning (ICML 1995), Morgan Kaufmann.

- Lewis, D. D. (1997). Reuters-21578 text categorization test collection. Distribution 1.0. from <http://www.daviddlewis.com/resources/testcollections/reuters21578>
- Liu, H.; & Motoda, H. (2007). *Computational Methods of Feature Selection*: Chapman & Hall/CRC.
- Mikolov, Tomas; Chen, Kai; Corrado, Greg; & Dean, Jeffrey. (2013). Efficient Estimation of Word Representations in Vector Space. *arxiv* pp 1-12. doi: 10.48550/arXiv.1301.3781
- Minaee, S; Kalchbrenner, N; Cambria, E; Nikzad, N.; Chenaghlu, M; & Gao, J. (2021). Deep Learning–Based Text Classification: A Comprehensive Review. *ACM Computing Surveys*, 54(3), pp. 1-40. doi: doi.org/10.1145/3439726
- Otunba, T; Lin, J. C. W; & Adewumi, A. O. (2022). Feature Selection Techniques for Text Classification: A Review. *Artificial Intelligence Review*, 55(6), pp. 4765-4804.
- Pennington, J; Socher, R; & Manning, C. D. (2014). *GloVe: Global vectors for word representation*. Paper presented at the Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)
- Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Comput. Surv*, 34(1), pp. 1-47
- Sun, Jian; Zhang, Xiang; Liao, Dan; & Chang, Victor. (2017). *Efficient method for Feature Selection in Text Classification*. Paper presented at the International Conference on Engineering and Technology (ICET).
- Uysal, A. K; & Gunal, S. (2014). The Impact of Preprocessing on Text Classification. *Information Processing & Management*, 50(1), pp. 104-112. doi: 10.1016/j.ipm.2013.08.006
- Yang; & Pederson, J. O. (1997). *A comparative study on feature selection in text categorization*. . Paper presented at the In Proceedings of the Fourteenth International Conference on Machine Learning.
- Yang, Yiming; & Liu, Xin. (1999). *A Re-examination of Text Categorization Methods*. Paper presented at the In Proceedings of the SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval, New York, NY, USA.