The Outcomes and Publication Standards of Research Descriptions in Document Classification: a Systematic Review Supplements

MARCIN MICHAŁ MIROŃCZUK* and ADAM MÜLLER*†, National Information Processing Institute, Poland
WITOLD PEDRYCZ‡, Department of Electrical & Computer Engineering University of Alberta, Canada

Document classification, a critical area of research, employs machine and deep learning methods to solve real-world problems. This study attempts to highlight the qualitative and quantitative outcomes of the literature review from a broad range of scopes, including machine and deep learning methods, as well as solutions based on nature, biological, or quantum physics-inspired methods. A rigorous synthesis was conducted using a systematic literature review of 102 papers published between 2003 and 2023. The 20 Newsgroups (bydate version) were used as a reference point of benchmarks to ensure fair comparisons of methods. Qualitative analysis revealed that recent studies utilize Graph Neural Networks (GNNs) combined with models based on the transformer architecture and propose end-to-end solutions. Quantitative analysis demonstrated state-of-the-art results, with accuracy, micro and macro F1-scores of 90.38%, 88.28%, and 89.38%, respectively. However, the reproducibility of many studies may need to be revised for the scientific community. The resulting overview covers a wide range of document classification methods and can contribute to a better understanding of this field. Additionally, the systematic review approach reduces systematic error, making it useful for researchers in the document classification community.

CCS Concepts: • **General and reference** → **Surveys and overviews**.

Additional Key Words and Phrases: document classification, text classification, systematic review, document classification review

ACM Reference Format:
Marcin Michał Mirończuk, Adam Müller, and Witold Pedrycz. 2023. The Outcomes and Publication Standards of Research Descriptions in Document Classification: a Systematic Review Supplements. *ACM Comput. Surv.* 0, 0, Article 0 (2023), 30 pages. https://doi.org/XXXXXXX.XXXXXXX

*Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Resources, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization
†Software, Formal analysis, Writing – Original Draft, Writing – Review & Editing, Visualization
‡Writing – Original Draft, Writing – Review & Editing

Authors' addresses: Marcin Michał Mirończuk, marcin.mironczuk@opi.org.pl; Adam Müller, Adam.muller@tutanota.com, National Information Processing Institute, al. Niepodległości 188 b, Warsaw, Masovian Voivodeship, Poland, 00-608; Witold Pedrycz, Department of Electrical & Computer Engineering University of Alberta, 116 St & 85 Ave, Edmonton, Canada, T6G 2R3, wpedrycz@ualberta.ca.

0360-0300/2023/0-ART0 $15.00
https://doi.org/XXXXXXX.XXXXXXX

1 COMPONENTS OF A TEXT CLASSIFICATION PIPELINE

The sections below outline the literature related to text classification extensively. More specifically, we deeply explain each related work to a text classification pipeline's different components. This analysis addresses questions 2-4 from our questionnaire.

1.1 Learning methods in the manipulation of input training data

Table 1 lists three articles that describe the use of training data.

Table 1. Known and used learning methods.

Aspect of work	Reference/Description
	Shen et al. [75]
Details	The authors propose applying learning vector quantization (LVQ) classifiers to the online scenario with stochastic gradient optimization for updating prototypes. They present two efficient clustering-based methods for extracting information from unlabeled data. They use different criteria to update prototypes for labelled and unlabeled data.
Findings	We can use both the maximum conditional likelihood criterion and the clustering criteria, such as Gaussian mixture or neural gas, alternatively based on the availability of label information. By doing so, we can fully leverage both supervised and unsupervised data to enhance performance.
Challenges	The authors failed to highlight any challenges or open problems.
	Kim et al. [45]
Details	The authors extend the standard co-training learning method. In order to increase the variety of feature sets for classification, they transform a document using three different document representation methods.
Findings	The proposed multi-co-training (MCT) method achieves superior classification performance, even when the documents are transformed into a very-low-dimensional vector and the labelled documents are very few.
Challenges	The work points out that, (1) different scenarios of class imbalance should be explored, and (2) the computational complexity should be improved.
	Pavlinek and Podgorelec [63]
Details	The authors propose a new self-training solution, known as Self-Training with Latent Dirichlet Allocation (ST-LDA), which utilizes inductive learning in which evaluation is conducted on a validation set.
Findings	Only a few instances in the early training stage caused the model to be overfitted. ST-LDA requires a minimal amount of training labelled data to outperform other models.
Challenges	The work points out that, (1) a better initial labeled set should be established, and (2) the computation time required should also be reduced.
	Cai and He [14]
Details	The authors propose an approach that explicitly considers the intrinsic manifold structure of data.
Findings	The authors suggest combining the methods that select the most uncertain data points, and those that select the most representative points.
Challenges	The work points out that, the computational complexity should be improved.

1.2 Pre-processing and feature construction

Table 2 lists one article that describes a pre-processing issue.

Table 2. Known and used pre-processing methods.

Aspect of work	Reference/Description
	Nagumothu et al. [60]
Details	The authors propose a hybrid approach that leverages the structure of knowledge embedded in a corpus. In particular, the paper reports on experiments where linked data triples (subject-predicate-object), constructed from natural language elements, are derived from deep learning.
Findings	The research indicates that linked data triples increased the F-score of the baseline GloVe representations by 6% and showed significant improvement over state-of-the-art models like BERT.
Challenges	The authors failed to highlight any challenges or open problems.

1.3 Feature weighting

Table 3 lists three articles that address the problem of feature weighting.

Table 3. Known and used schemes of feature weighting.

Aspect of work	Reference/Description
	Attieh and Tekli [6]
Details	The authors present a new text classification framework called Category-based Feature Engineering (CFE). It includes a supervised weighting scheme based on a variant of the Term Frequency-Inverse Category Frequency (TF-ICF) model, integrated into three efficient classification methods.
Findings	The proposed approach improves text classification accuracy while requiring significantly less computation time than their deep model alternatives.
Challenges	The paper suggests three areas for further research, (1) exploring the use of external corpora and semantic data augmentation to enhance target feature vectors, (2) utilizing human-tailored knowledge bases such as WordNet and DBPedia, and (3) conducting more comprehensive evaluation procedures.
	Shehzad et al. [74]
Details	The authors propose a novel approach for term weighting called the binned term count (BTC), which involves a non-linear mapping of term frequency.
Findings	BTC helps to mitigate the normalization effect on lengthy documents.
Challenges	The study recommends using variable bin lengths that adjust to the average size of documents in a corpus. This approach may control and allocate document length variations more effectively.
	Jia and Zhang [36]
Details	A new term weighting scheme called Document Representation Based on Global Policy (DRGP) is introduced.
Findings	We should choose the representation methods according to corpora characteristics for better classification performance.
Challenges	The study recommends continuing the research and introducing new optimization methods to reduce the calculation cost.

	Tang et al. [81]
Details	A new term weighting scheme called Term frequency & Inverse gravity moment (TF-IGM) and its variants are introduced.
Findings	The paper demonstrates that TF-IGM performs better than TF-IDF and current supervised term weighting methods. Furthermore, the study presents new and thoroughly analyzed findings that differ from previous research.
Challenges	The study recommends (1) conducting comparative studies on the IGM model as a new measure of sample distribution non-uniformity and the traditional statistical models such as variance and entropy, (2) expanding the scope of experiments for text classification, and (3) applying the IGM model to feature dimension reduction and sentiment analysis.
	Wang et al. [89]
Details	A new term weighting entropy-based schemes to measure the effectiveness of terms in distinguishing between categories are introduced.
Findings	Term weighting scheme effectiveness varies with datasets, classifiers, and classification types. The authors propose their schemes as better reflecting term distinguishing power in text categorization than many previous schemes.
Challenges	The study recommends (1) evaluating the method on larger datasets, (2) improving model parameter estimation, and (3) exploring the potential benefits of incorporating entropy-based term weighting methods to enhance the performance of embedding methods.
	Tang et al. [80]
Details	A new term weighting scheme called Frequency-inverse Exponential Frequency (TF-IEF), with a new global weighting factor, IEF, to characterize a global weighting factor is introduced.
Findings	TF-IEF outperforms other term weighting schemes, such as TF-CHI2 and TF-IG.
Challenges	The authors failed to highlight any challenges or open problems.
	Chen et al. [17]
Details	A new Supervised Term Weighting (STW) scheme called Term Frequency & Inverse Gravity Moment (TF-IGM) is introduced.
Findings	TF-IGM outperforms TF-IDF and the state-of-the-art STW schemes.
Challenges	The work points out that, (1) comparative studies on the IGM model as a new measure of sample distribution should be conducted, and (2) the model should be applied to feature dimension reduction and sentiment analysis.
	Luo et al. [58]
Details	The authors propose a novel term weighting scheme by exploiting the semantics of categories and indexing terms.
Findings	The approach outperforms TF-IDF with small amounts of training data, or when the content of the documents is focused on well-defined categories.
Challenges	The work points out that, (1) other ontologies, with wider coverage for expressing the sense of words and category labels, should be employed, and (2) different ways of representing the semantics of categories and other similarity measures should be explored.

1.4 Learning algorithms that utilize dimension reduction techniques - feature selection case

Table 4 lists fifteen articles that propose new feature selection methods.

Table 4. Known and used feature selection methods.

Aspect of work	Reference/Description
	Brockmeier et al. [13]
Details	The feature selection proposed utilizes a method known as descriptive clustering, which involves automatically organizing data instances into clusters, and generating a descriptive summary for each cluster.
Findings	The proposed method performs accurately, and yields feature subsets that are indicative of the cluster content.
Challenges	The work points out that, the more complex features, including multi-word expressions, named entities, and clusters of the features themselves should be investigated.
	Al-Salemi et al. [2]
Details	Seven feature ranking methods are applied in order to improve the performance of the RFBoost classification method [3]. An accelerated version of RFBoost, called RFBoost1, is introduced.
Findings	RFBoost is an improved and accelerated version of AdaBoost. MH. There is no overall best feature ranking method. The performance of the feature ranking methods depends on the nature of the datasets. RFBoost1 has fast performance.
Challenges	The authors wish to investigate the use of other feature selection methods to improve both RFBoost and RFBoost1.
	Hassaine et al. [32]
Details	The proposed approach extracts keywords in hierarchical order of importance using a hyper rectangle tree.
Findings	The hyper rectangle algorithm provides discriminating features that are almost independent of the chosen weights. The logistic regression classifier outperforms random forests due to better handling of a large number of features.
Challenges	The work points out that, the other tasks, such as anomaly detection, sentiment analysis, and document indexing and ranking should be considered.
	Javed and Babri [35]
Details	A method known as Normalized Difference Measure (NDM) utilizes the true positive rate (tpr) and false positive rate (fpr) to create a feature ranking metric.
Findings	A term occurring with different document frequencies in positive and negative classes is relatively more discriminative than one that has similar document frequencies in both classes. NDM boosts terms that are rare in both classes.
Challenges	The authors failed to highlight any challenges or open problems.
	Tang et al. [78]
Details	The authors use Baggenstoss's PDF Project Theorem (PPT) to reformulate Bayes' decision rule for classification with selected class-specific features.
Findings	An improvement is achieved in a small number of the features. When more features are selected, the classification performance of all methods considered improve, leading to a minor improvement to the overall approach.
Challenges	The authors failed to highlight any challenges or open problems.

	Tang et al. [79]
Details	Based on JMH-divergence, the authors developed two efficient feature selection methods for text categorization, termed maximum discrimination (MD) and χ^2 methods.
Findings	Almost all of the filter-based feature selection approaches use binary variables. These filter approaches only by exploring the intrinsic characteristics of the data.
Challenges	The work points out that, (1) feature dependence should be utilized in order to maximize discriminative capacity, and (2) enhancement of the learning for rare categories should be considered.
	Li [48]
Details	The authors propose a formula that convert the values of a feature selection method's parameters from integers to real values.
Findings	The proposed method assists the Chi-square (χ^2) and Information Gain (IG) metrics to obtain better results, especially when fewer features are used on imbalanced datasets.
Challenges	The work points out that, (1) other datasets should be considered for evaluation, and (2) the proposed strategy should be applied to revising other feature selection methods.
	Al-Salemi et al. [3]
Details	The RFBoost algorithm proposed is based on filtering a low, fixed number of ranked features, rather than using all features. Two methods for ranking features are proposed: (1) One Boosting Round (OBR); and (2) Labeled Latent Dirichlet Allocation (LLDA).
Findings	RFBoost, with the new weighting policies and the LLDA-based feature ranking, significantly outperformed all other algorithms evaluated. OBR-based feature ranking yielded the worst performance overall.
Challenges	The work points out that, (1) multi-label classification problems should be considered, and (2) other feature ranking methods, as it is the core concept for improving the effectiveness of RFBoost, should be considered.
	Wang et al. [87]
Details	The proposed method, known as Categorical Document Frequency Divided by Categorical Number (CDFDC), involves adding information about categories to a given term in the original formula of Categorical Document Frequency (CDF), to increase the discrimination of the terms.
Findings	The high computational complexity might not enable high precision or recall rate, and stable and predictable time efficiency is necessary.
Challenges	The work points out that, (1) other, more extensive datasets should be evaluated, and (2) the connection between algorithm complexity and document-time-efficiency should be explored.
	Zong et al. [108]
Details	The method established, known as Discriminative Feature Selection with Similarity (DFS+Similarity), selects features with strong discriminative power, and considers the semantic similarity between features and documents.

Findings	DFS, DFS+Similarity, Chi Square ($\chi 2$) statistic, Information Gain (IG), and Mutual Information (MI) produce the worst results when the number of features is the lowest (1000). MI and IG perform relatively poorer than the others, and they are sensitive to the number of features.
Challenges	The work points out that, (1) multi-label classification problems should be considered, and (2) the characteristics of feature distribution in each category of documents should be considered.
	Feng et al. [23]
Details	The authors develop an optimal set of features that is characterized by global and local section indices for group and single features, respectively. The author also proposes a Latent Selection Augmented Naive (LSAN) Bayes classifier to enable a suitable fit to the data.
Findings	Feature selection and feature weighting can be combined organically in the classifier proposed. The high dimension can be reduced in this model when working with not only the feature selection indices, but also future predictions.
Challenges	The work points out that, (1) other feature selection and weighting methods and parameters should be examined, (2) corresponding laws for feature selection should be explored, and (3) a statistically in-depth analysis should be performed.
	Li et al. [50]
Details	The method, known as Weighted Document Frequency (WDF), creates a feature ranking based on information about how a feature is essential for a document.
Findings	The method outperforms the document frequency (DF) approach, but there is no difference between the proposed method and the Chi-Square ($\chi 2$) method.
Challenges	The work points out that, (1) more research with other datasets is needed, (2) other text mining applications should be considered for evaluation, and (3) the influence of feature-weighting schema should be studied more deeply.
	Rehman et al. [67]
Details	A new feature ranking metric called Relative Discrimination Criterion (RDC) enhances the ranking of the terms present in only one class, or those for which the term counts in a single class are relatively higher than in the other.
Findings	The method selects suitable features. The RDC measure, however, requires somewhat more computation than the other feature ranking metrics.
Challenges	The work points out that, (1) more research with other datasets is needed, and (2) tuning of the parameters should be considered.
	Yan et al. [96]
Details	The proposed optimization framework integrates feature selection and feature extraction. A novel feature selection algorithm called Trace Oriented Feature Analysis (TOFA) optimizes the feature extraction objective function in the solution spaces of feature selection algorithms.
Findings	Many commonly used algorithms are special cases of the proposed unified objective function. The optimal solution, according to its objective function, can be achieved. TOFA is suitable for large-scale text data. The solution can handle both unsupervised and semi-supervised cases.
Challenges	The work points out that, (1) the relationship between data distribution and the framework's parameters should be established, and (2) calibration of the optimal setting should be performed.
	Tesar et al. [82]

Details	A new suffix-tree-based algorithm discovers itemsets which contains different and valuable words.
Findings	Bigrams seem to be more suitable for text classification. A feature subset selection approach should be determined on the basis of the principle of the classifier used. The extended bag-of-words (BoW) failed to overcome the best results achieved by the simple BoW approach.
Challenges	The work points out that, a more in-depth experiment and evaluation should be performed to establish what types of activity influence text classification performance significantly.

1.5 Learning algorithms that utilize dimension reduction techniques - feature projection case

Table 5 lists twelve articles that address the problem of feature projection.

Table 5. Known and used feature projection methods.

Aspect of work	Reference/Description
	Guo and Yao [30]
Details	The authors propose the concept of containers and further explore the properties of word containers and document containers through experiments and theoretical demonstrations.
Findings	The document container has a fixed capacity, and the document vector obtained by a simple average of too many word embeddings cannot be fully loaded by the container. It will lose some semantic and syntactic information on vast text datasets.
Challenges	The authors failed to highlight any challenges or open problems.
	Jiang et al. [37]
Details	The authors introduce the task of conceptual labelling, which aims to generate the minimum number of concepts as labels to represent and explicitly explain the semantics of a Weighted Bag of Words (WBoW).
Findings	Experiments and results prove that the proposed method can generate proper labels for WBoWs.
Challenges	The authors' future work focuses on properly incorporating conceptual labels into some NLP tasks.
	Unnam and Reddy [84]
Details	The authors propose a framework to represent a document in a unique feature space. They do this by assigning each dimension a potential feature word with high discriminatory power. The model then computes the distances between the document and the feature words.
Findings	The proposed model outperforms baseline methods in document classification and uses interpretable word features to represent the document. It offers an alternative framework for representing larger text units with word embeddings and provides opportunities for developing new approaches to improve document representation and its applications.

Challenges	The study recommends (1) extending the proposed model to enhance its performance, (2) combining the selection criteria for a hybrid feature word selection approach, and (3) developing a word weighting scheme that uses frequency and radius to improve performance.
	Yang et al. [100]
Details	The authors provide a new Graph Attention Topic Network (GATON) method to overcome the overfitting issue of Probabilistic Latent Semantic Indexing (pLSI).
Findings	The GATON model is designed to capture the topic structure of documents. This is achieved through the use of graph neural networks, which are equivalent to semi-amortized inference of stochastic block model (SBM) on network data. Similarly, pLSI is equivalent to SBM on a specific bi-partite graph.
Challenges	The authors failed to highlight any challenges or open problems.
	Białas et al. [10]
Details	The authors propose a novel biologically plausible mechanism for generating low-dimensional spike-based text representation.
Findings	It is recommended that inhibition be disabled during the (Spiking Neural Network) SNN evaluation phase. Pruning out as many as 90% of connections with the lowest weights did not affect the representation quality while heavily reducing the SNN computational complexity, i.e. the number of differential equations describing the network.
Challenges	The work points out that we should explore the opportunity to expand the SNN encoder towards Deep SNN architecture by adding more layers of spiking neurons, allowing us to learn more detailed features of the input data.
	Chen and Feng [15]
Details	The authors propose a novel Boltzmann bases feature extraction called Gaussian Fuzzy Restricted Boltzmann Machine (GFRBM) for real-valued inputs.
Findings	The authors found that the proposed solution outperforms discriminative RBM models regarding reconstruction and classification accuracy. They behave more stably when encountering noisy data.
Challenges	The work suggests that a more efficient learning algorithm and deep fuzzy models based on FRMB variants should be developed.
	Kesiraju et al. [43]
Details	The authors present the Bayesian subspace multinomial model (Bayesian SMM). This generative log-linear model learns to represent documents in the form of Gaussian distributions, thereby encoding the uncertainty in its covariance.
Findings	The perplexity measure valuation shows that the proposed Bayesian SMM fits the unseen test data better than the state-of-the-art neural variational document models. Also, the proposed systems are robust to over-fitting unseen test data.
Challenges	The work points out that other scoring mechanisms that exploit the uncertainty in embeddings should be explored.
	Li et al. [52]
Details	A proposed representation scheme known as Bag-of-Concepts (BoC) automatically acquires useful conceptual knowledge from an external knowledge base. A second representation model, known as Bag-of-Concept-Clusters (BoCCl), improves BoC representation further.

Findings	Bag-of-words (BoW) is a solid baseline for document classification tasks. BoC and BoCCl can effectively capture the concept-level information of documents. They also offer high interpretability.
Challenges	The work points out that, (1) sentence-level-based representation should be considered, and (2) conceptual knowledge should be incorporated into deep neural networks.
	Gupta et al. [31]
Details	The authors propose an extension of Sparse Composite Document Vector (SCDV) called SCDV-MS utilizes multi-sense word embeddings and learns a lower dimensional manifold.
Findings	The SCDV-MS embeddings proposed in the study are more efficient than SCDV regarding time and space complexity for textual classification tasks. Disambiguating multi-sense words using adjacent words in the context can result in improved document representations. The representation noise at the word level can significantly affect downstream tasks.
Challenges	The authors failed to highlight any challenges or open problems.
	Yang et al. [99]
Details	An innovative latent relation-enhanced word embedding model increases the semantic relatedness of words in the corpus. The authors discover more useful relations between words, and add them to word embeddings.
Findings	Word embedding representation is a powerful tool, as it served various systems as a reliable input.
Challenges	The work points out that, the contextual information, as the unique distributions to generate word embeddings should be analyzed carefully.
	Chen and Zaki [19]
Details	Competitive learning is introduced during an autoencoder training phase. Due to the competition between neurons, each becomes specialized, and the overall model can learn meaningful representations of textual data.
Findings	The proposed autoencoder, known as KATE, can learn better representation than traditional autoencoders, and outperforms deep generative models, probabilistic topic models, and even word representation models.
Challenges	The work points out that, (1) KATE should be evaluated on more domain-specific datasets, (2) the scalability and effectiveness of the approach should be improved.
	Hu et al. [34]
Details	The authors developed a new regularized Restricted Boltzmann Machines (RBMs), which accounts for class information.
Findings	The features extracted by the proposed method have strong discriminant power. The improved performance of the method comes at the cost of high computational demands. The effect of the inter-class repulsion regularization component obtained by the models is imperceptible — features of different groups cannot be effectively separated.
Challenges	The work points out that, a new inter-class repulsion regularization should be used to improve the performance of the method.
	Kesiraju et al. [42]
Details	A Subspace Multinomial Model (SMM), in which modification, i.e. regularization of terms creates a compact and continuous representation for the documents.

Findings	The classification accuracy of the SMM increases with the dimensionality of the latent variable, which is not the case with Sparse Topical Coding (STC) or Probabilistic Topic Models (PTM).
Challenges	The work points out that, (1) an in-depth exploration of different optimization techniques should be performed, and (2)this should involve exploring discriminative SMMs, and fully Bayesian modelling of SMMs.
	Li et al. [51]
Details	A novel hybrid model known as Mixed Word Embedding (MWE) combines two variants of Word2Vec seamlessly by sharing a common encoding structure. Moreover, the model incorporates a global text vector in order to capture more semantic information.
Findings	MWE achieves highly competitive performance. MWE preserves the same time complexity as the Skip-Gram model.
Challenges	The work points out that, MWE should be improved by incorporating more external corpora, and giving consideration to proximity and ambiguity among words.
	Zheng et al. [104]
Details	A Bidirectional Hierarchical Skip-Gram model (BHSG), models text topic embedding, and considers a whole sentence or document as a special word to capture the semantic relationship between the words and the global context word.
Findings	BHSG utilizes negative sampling; thus, it is highly suitable for large scale data.
Challenges	The work points out that, BHSG should be extended to implement more topic-related tasks, such as keyword extraction and text summarization.
	Rodrigues and Engel [69]
Details	The proposed method is based on the Incremental Naive Bayes Clustering (INBC) algorithm, which was initially designed for continuous inputs, and so is considered an extension of it.
Findings	A single pass over the training data is required to achieve an impressive classification result. As more data is presented, the model can be improved.
Challenges	The work points out that, (1) feature selection should be performed, and (2) other properties, such as similarity criteria, should be verified.
	Cai and He [14]
	See Table 1
	Li et al. [53]
Details	A Concise Semantic Analysis (CSA) technique extracts a few concepts (a new N-dimensional space, in which each concept represents a new dimension) based on class labels. It then implements a concise interpretation of words and documents in this new space.
Findings	The CSA helps with dimension-sensitive learning algorithms, such as k-nearest neighbors, to eliminate the *Curse of Dimensionality* [1, 11]. K-nearest neighbors in the new concept space performs comparably with SVMs. CSA performs equally well in the Chinese and English languages, and incurs a very low computational cost.
Challenges	The work points out that, (1) CSA should adopted to perform on large taxonomies of text categorization, and (2) the technique should be also parallelized.
	Salakhutdinov and Hinton [71]

Details	The authors propose a method that creates a separate Restricted Boltzmann Machines (RBM) for each document, with as many Softmax units as there are words in the document. The authors also present efficient learning and inference algorithms for the model.
Findings	The model's learning is easy and stable. It may be scaled up to classify billions of documents. This is in contrast to directed topic models, in which most of the existing inference algorithms are designed to be run in a batch mode. The proposed model can generalize much better than Latent Dirichlet Allocation (LDA).
Challenges	The work points out that, (1) label information should be added to the modelling, (2) the document-specific metadata observed should be incorporated into the model's learning, and (3) more layers should be added to create a Deep Belief Network [33].
	Yan et al. [96]
	See Table 4

1.6 Learning algorithms with new classification methods

Table 6 lists 27 articles that propose new classification methods.

Table 6. Known and used classification methods.

Aspect of work	Reference/Description
	Wang et al. [90]
Details	The authors propose a novel Text Classification by Fusing Contextual Information via Graph Neural Networks (TextFCG) that fuses contextual information and handles documents with new words and relations.
Findings	Text-FCG outperforms other methods for short- and medium-length text, while sparse graphs or topic models are more effective for long texts. Compared to other graph-based models, Text-FCG shows significant improvements, underscoring the importance of diverse contextual information in learning text representations.
Challenges	The authors failed to highlight any challenges or open problems.
	Khandve et al. [44]
Details	The authors explore hierarchical transfer learning approaches for long document classification. In a hierarchical setup, they employ pre-trained Universal Sentence Encoder (USE) and Bidirectional Encoder Representations from Transformers (BERT) to capture better representations efficiently.
Findings	The USE with CNN/LSTM performs better than its stand-alone baseline. In contrast, the BERT with CNN/LSTM performs on par with its stand-alone counterpart.
Challenges	The authors failed to highlight any challenges or open problems.
	Guidotti and Ferrara [28]
Details	The authors regard the text as a superposition of words, derive a document's wave function and compute the document's transition probability to a target class according to Born's rule.

Findings	The proposed classifier is self-explainable and can be embedded in neural network architectures. Also, the results suggest that physical principles can be successfully exploited in machine learning and may open a new class of classification algorithms.
Challenges	The paper suggests several potential improvements and extensions of the work: (1) the effectiveness of the method in the view of machine learning remains an open question, (2) the construction of a wave function explicitly should be considered, and (3) the construction of deep networks that apply the transformation should be developed.
	Yang et al. [98]
Details	The paper introduces a new type of neural network called Simple Jumping Knowledge Networks (SJK-Nets), a two-step process. First, a simple no-learning method completes the neighbourhood aggregation process. Then, a ”jumping architecture“ combines each node’s different neighbourhood ranges to represent the network structure better.
Findings	The authors highlight that - SJK-Nets’ neighbourhood aggregation is a no-learning process, so SJK-Nets are successfully extended to node clustering tasks.
Challenges	The paper suggests three possible future research directions: (1) exploring other layer aggregators, (2) extending the method to more downstream tasks, and (3) applying it to other fields beyond graph-based tasks such as natural language processing and computer vision.
	Dai et al. [21]
Details	The authors propose a Graph Fusion Network (GFN) to enhance text classification performance. It builds homogeneous text graphs with word nodes in the graph construction stage and transforms external knowledge into structural information. The graph reasoning stage involves graph learning, convolution, and fusion, where a multi-head fusion module integrates different opinions. GFN can make inferences on new documents without rebuilding the whole text graph.
Findings	Experimental results demonstrate the method’s superiority. Notably, the diverse graph views are mutually beneficial. The well-crafted multi-head fusion module effectively enhances the system’s performance.
Challenges	The paper suggests (1) a thorough investigation into deep learning models’ interpretability, (2) exploring alternative methods to construct text graphs, and (3) conducting a comparative analysis between BERT-style and GNN-based models.
	Prabhakar et al. [64]
Details	The authors describe two new techniques for improving deep learning models. The first is called the Evolutionary Contiguous Convolutional Neural Network (ECCNN), where the data instances of the input point are considered along with the contiguous data points in the dataset. The second technique is Swarm DNN, a swarm-based Deep Neural Network that utilizes Particle Swarm Optimization (PSO) for text classification.
Findings	The two proposed models achieved satisfying results.
Challenges	Future works aim to work with many other nature-inspired and ensemble deep-learning models for efficient text classification.

	Zhu and Koniusz [107]
Details	The authors propose to use a modified Markov Diffusion Kernel to derive a variant of Graph Convolution Networks (GCNs) called Simple Spectral Graph Convolution (S2GC).
Findings	The S2GC method utilizes low- and high-pass filters to capture the global and local contexts of each node. This technique outperforms competitors by effectively aggregating over more prominent neighbourhoods while avoiding over-smoothing.
Challenges	The authors failed to highlight any challenges or open problems.
	Lin et al. [54]
Details	The authors propose BertGCN, a model combining large-scale pretraining and transductive learning for text classification. The model propagates label influence through graph convolution to learn representations for training and unlabeled test data.
Findings	BertGCN achieves state-of-the-art performance on various text classification datasets, as demonstrated in experiments. The framework can be built on any document encoder and graph model.
Challenges	The authors suggest that using document statistics to construct the graph may not be as optimal as models that can automatically create edges between nodes. Therefore, addressing this issue could be a possible feature direction.
	Yan et al. [97]
Details	The article proposes a model that combines quantum probability and Graph Neural Networks (GNNs) to capture the global structure of interactions between documents for document representation and classification.
Findings	The comprehensive analyses illustrate the proposed model's resilience to limited training data and its capability to learn semantically distinct document representations.
Challenges	The authors failed to highlight any challenges or open problems.
	Wang et al. [88]
Details	The authors outperform Bi-filtering Graph Convolutional Network (BGCN) thanks to simply cascading two sub-filtering modules. The new solution is called Simple Bi-filtering Graph Convolution (SBGC) framework and is inspired by the direct implementation of Infinite Impulse Response (IIR) graph filters.
Findings	Experiments show that SBGC outperforms other methods in performance and computational efficiency and that BGCN and SBGC are robust to feature noise and exhibit high label efficiency.
Challenges	The paper suggests two possible future research directions: (1) developing filters with the flexibility of IIR filters for different scenarios and (2) reconsidering the design principles of graph convolution networks based on graph signal processing.
	Ragesh et al. [66]
Details	The authors introduce HeteGCN, a novel heterogeneous graph convolutional network model that leverages the predictive text embedding (PTE) and TextGCN approaches.
Findings	Reducing model parameters can improve training speed and performance in scenarios with limited labelled data.

Challenges	The paper proposes three future research directions: (1) investigating the advantages of the HeteGCN-BERT augmented model, (2) expanding the model to address recommendation problems, and (3) integrating knowledge graphs into the model.
	Xie et al. [94]
Details	The paper introduces the Graph Topic Neural Network (GTNN), a novel model capable of learning latent topic semantics and generating an interpretable document representation by accounting for relationships among documents, words, and the graph structure.
Findings	The model can embed richer structural semantics in the learned representation for downstream tasks. Furthermore, it addresses the well-known interpretability problem of the learned document representations in previous GCN-based methods.
Challenges	The work indicates that we should extend the model to large-scale datasets and online learning. Moreover, we should also incorporate linguistic resources such as Wordnet into the graph.
	Xie et al. [95]
Details	The authors describe a proposed model called Topic Variational Graph Auto-Encoder (T-VGAE) that combines a topic model with a variational graph-auto-encoder to capture hidden semantic information between documents and words.
Findings	The proposed method is more interpretable than similar methods and can deal with unseen documents.
Challenges	The work points out that exploring better-suited prior distribution in the generative process would be interesting. Extending the model to other tasks, such as information recommendation and link prediction, is also possible.
	Zhou et al. [105]
Details	The authors suggest two modules to address the limitations of standard Convolutional Neural Networks (CNNs): one utilizes discriminative filters (filters with a maximum divergence). At the same time, the other enables the complete extraction of all essential features.
Findings	The proposed model increases the discriminative power of the model by maximizing the distance between different filters and a novel global pooling mechanism for feature extraction.
Challenges	The authors' future work will focus on adequately incorporating conceptual labels into some NLP tasks.
	Zhang and Yamana [102]
Details	The authors discuss an alternative approach to encoding labels into numerical values by incorporating label knowledge directly into the model without changing its architecture.
Findings	The experimental results demonstrate that the proposed method can understand the relationship between sequences and labels.
Challenges	Future work in this area includes two key directions: (1) developing an appropriate keyword set for representing label knowledge without introducing noise and (2) identifying improved methods for calculating relatedness beyond simply using the mean value.
	Liu et al. [55]

Details	TensorGCN (tensor graph convolutional networks) is a new framework that uses a text graph tensor to capture semantic, syntactic, and sequential contextual information. Intra-graph and inter-graph propagation learning are used to aggregate information from neighbouring nodes and harmonize heterogeneous information between graphs.
Findings	The proposed TensorGCN presents an effective way to harmonize and integrate heterogeneous information from different graphs. The inter-graph propagation strategy is crucial for graph tensor learning.
Challenges	The authors failed to highlight any challenges or open problems.
	Wei et al. [92]
Details	The proposed model utilizes a recurrent structure to retain word order and capture contextual information, and incorporates message passing from the graph neural networks (GNNs) to update word hidden representations. Additionally, a max-pooling layer is used to capture critical components in text for classification, similar to GNN's readout operation.
Findings	The experimental results show that the model significantly improves against the RNN-based and GNN-based models. The model is suitable for constructing the semantic representation of the entire text. The performance of Text GCN highly depends on the quality of text graph, which limits its scope of application.
Challenges	The paper suggests that (1) long-distance contextual information may be lost, (2) additional layers could improve the model's performance, and (3) employing a recurrent structure would enhance the capture of long document contextual information.
	Chiu et al. [20]
Details	The authors' model uses an attention mechanism to dynamically decide how much information to use from a sequence - or graph-level component.
Findings	The proposed graph-level extensions enhance performance on most benchmarks. Furthermore, the adapted attention-based architecture outperforms the generic and fixed-value concatenation alternatives. These extensions for text classification enable the system to learn diverse inter-sentential patterns.
Challenges	The authors fail to highlight any challenges or open problems.
	Wang et al. [91]
Details	The authors propose a new trainable hierarchical topic graph (HTG) incorporating a probabilistic deep topic model into graph construction. The HTG consists of word-level, hierarchical topic-level, and document-level nodes, which exhibit a range of semantic variations from fine-grained to coarse.
Findings	The proposed model, called dynamic HTG (DHTG), uses Graph Convolutional Networks (GCN) for variational inference to evolve the HTG for end-to-end document classification dynamically. The model can also learn an interpretable document graph with meaningful node embeddings and semantic edges.
Challenges	The authors fail to highlight any challenges or open problems.
	Ding et al. [22]
Details	The authors utilize document-level hypergraphs to model text documents and introduce a new group of GNN models called HyperGAT for generating distinctive text representations.
Findings	The proposed model is (1) unable to capture high-order interaction between words and (2) inefficiently handling large datasets and new documents.

Challenges	The authors fail to highlight any challenges or open problems.
	Zhou et al. [106]
Details	The authors propose a Discriminative Convolutional Neural Network with Context-aware Attention to solve the challenges of vanilla Convolutional Neural Networks (CNN). The proposed solution encourages discrimination across different filters via maximizing their earth mover distances and estimates the salience of feature candidates by considering the relation between context features.
Findings	The proposed model can capture representative semantics and effectively compute feature salience for a specific task.
Challenges	The paper suggests two key points, (1) adopting an adaptive method to extract valuable features of flexible sizes using a context-aware mechanism, and (2) applying the model to related tasks like relation classification and event extraction.
	Guo and Yao [29]
Details	The work aims to create a valuable and effective word and document matrix representation architecture based on a linear operation to learn representations for document-level classification.
Findings	A convolutional-based classifier is more suitable for the document matrix. The convolution operation can better capture the proposed document matrix's two-dimensional features by analysing theoretical and experimental perspectives.
Challenges	The authors fail to highlight any challenges or open problems.
	Chen and Srihari [16]
Details	The authors are interested in deep learning for classification with prior, where the labels are expressed in a hierarchy. In particular, they attempt to leverage knowledge transfer and parameter sharing among classes.
Findings	The proposed model shows promising results compared to support vector machines and other deep learning methods. Also, the model inherits the advantages of deep learning and can handle overfitting and reduce the redundancy between node parameters.
Challenges	The work points out that transfer learning should be considered. Another topic deserving of exploring is how to learn the structural prior.
	Aler et al. [4]
Details	The main aim of this article is to carry out an extensive investigation on essential aspects of using Hellinger Distance (HD) in Random Forests (RF), including handling multi-class problems, hyper-parameter optimization, metrics comparison, probability estimation, and metrics combination.
Findings	The results demonstrate HD's robustness in RF, but it has some limitations for balanced multi-class datasets. Combining metrics can enhance performance. Nevertheless, Gini appears to be more suitable than HD when applied to text datasets.
Challenges	HD is inferior to Gini for text classification, making it crucial to investigate the underlying reasons. Additionally, considering other distribution distances like Kullback-Leibler divergence as substitutes to HD is worth exploring.
	Yao et al. [101]

Details	The authors build a single text graph for a corpus based on word co-occurrence and document word relations, then learn a Text Graph Convolutional Network (Text GCN) for the corpus.
Findings	The proposed Text GCN method excels in text classification, surpassing state-of-the-art approaches and acquiring predictive word and document embeddings. It also demonstrates robustness to less training data, further highlighting its effectiveness.
Challenges	The work points out that we should improve the classification performance using attention mechanisms and develop an unsupervised text GCN framework for representation learning on largescale unlabeled text data.
	Tiwari and Melucci [83]
Details	A novel classification method known as a Quantum-Inspired Binary Classifier (QIBC) resolves a binary classification problem. It is inspired by quantum detection theory.
Findings	QIBC can outperform the baselines in several categories. Some results, however, remain unsatisfactory in some categories.
Challenges	The work points out that, (1) an in-depth error classification analysis should be performed, and (2) multi-label classification problems should be addressed.
	Berge et al. [9]
Details	A Tsetlin Machine learns propositional formulae, such as IF “rash” AND “reaction” AND “penicillin” THEN Allergy, to represent the particular facets of each category.
Findings	The proposed method captures categories using simple propositional formulae that are readable to humans. The explanatory power of Tsetlin Machine-produced clauses seems to equal that of decision trees.
Challenges	The work points out that, (1) a utilization of word embeddings should be considered, (2) a combination of different data views should be applied, and (3) datasets with more complicated structures should be considered.
	Unnikrishnan et al. [85]
Details	This article proposes a new approach to sparse classification, and presents a comparative study of different sparse classification strategies for text classification.
Findings	The minimum reconstruction error criterion is suitable for the problem of text classification. The computational bottle-neck can be resolved using the proposed dictionary refinement procedure.
Challenges	The authors failed to highlight any challenges or open problems.
	Pappagari et al. [62]
Details	A new multi-scale Convolutional Neural Network (CNN) architecture that uses raw text as input. It contains parallel convolutional layers, and jointly optimises a new objective function, which, in turn, optimizes two tasks simultaneously.
Findings	The objective function, which integrates the verification and identification tasks, improves the results of the identification tasks. This approach does not use text pre-processing to achieve better document classification performance.
Challenges	The work points out that, the sequence dynamics modeling with Long Short-Term Memory (LSTM) should be incorporated into the proposed model.
	Al-Salemi et al. [2]
	See Table 4

	Feng et al. [24]
Details	The authors consider the overfitting problem and propose a quantitative measurement, rate of overfitting, denoted as RO. They also propose an algorithm known as AdaBELM.
Findings	Extreme Learning Machines (ELMs) suffer from a significant overfitting problem. The proposed model, AdaBELM, resolves this drawback and has high generalizability, which is demonstrated by its high performance.
Challenges	The authors failed to highlight any challenges or open problems.
	Sharma et al. [72]
Details	The article proposes a new hierarchical sparse-based classifier, exploring the concept of sparse coding for text classification, and seeding the dictionary used the principal components.
Findings	The proposed hierarchical classifier works better than flat sparse-based classifiers. Principal Component Analysis (PCA) may be used to create an overcomplete dictionary.
Challenges	The work points out that, (1) more research with other datasets should be conducted, and (2) the semantic information should be considered.
	Benites and Sapozhnikova [8]
Details	The work explores a scalable extension—a Hierarchical Adaptive Resonance Associative Map (HARAM)—to a fuzzy Adaptive Resonance Associative Map (ARAM) neural network for quick classification of high-dimensional and large data.
Findings	HARAM is faster than ARAM. A voting classification procedure increases its accuracy. Adaptive Resonance Theory (ART) neural networks are highly parallelized.
Challenges	The authors noted that the details of implementation could be an issue.
	Johnson and Zhang [41]
Details	The work aims to create a valuable classification method of documents under the one-hot CNN (convolutional neural network) framework. The authors explore a more sophisticated region embedding method using Long Short-Term Memory (LSTM).
Findings	The study shows that embeddings of text regions, which can convey complex concepts, are more valuable than embeddings of single words in isolation.
Challenges	A promising future direction might be to seek, under this framework, new region-embedding methods with complementary benefits.
	Sharma et al. [73]
Details	The work explores the idea of sparse coding for text classification and seeding the dictionary using principal components. The article also explores the use of Support Vector Machines (SVMs) with frequency-based kernels.
Findings	PCA may be utilized to create an overcomplete dictionary. SVMs with Hellinger's kernel, and without PCA, produces the best results. A voting classification procedure improves the outcomes.
Challenges	The work points out that, (1) the semantic information should be taken into account, and (2) better strategies for combining the classifiers must be explored.
	Jin et al. [38]
Details	The authors built text classifier by using a Naive Bayes model, utilizing a new structure called bag-of-embeddings probabilities.

Findings	The model is conceptually simple; the only parameters being embedding vectors, trained using a variation of the Skip-gram method. The proposed model outperforms state-of-the-art methods for both balanced and imbalanced data.
Challenges	The work points out that, (1) leveraging unlabeled data for semi-supervised learning should be considered, and (2) other neural document models should be exploited to achieve higher accuracy.
	Al-Salemi et al. [3]
	See Table 4
	Pang et al. [61]
Details	A new classification method called CenKNN combines the strengths of two widely-used text classification techniques, k-nearest neighbors and Centroid.
Findings	CenKNN overcomes the drawbacks of k-nearest neighbors classifiers. CenKNN works better than Centroid. The proposed method is appropriate for highly imbalanced corpora with a low number of classes. SVM is a better choice for large balanced corpora.
Challenges	The work points out that, (1) CenKNN should be improved to handle subclusters and/or a larger number of classes, and (2) multi-label classification problems should be addressed.
	Kusner et al. [46]
Details	A distance function, Word Mover's Distance (WMD) measures the dissimilarity between two text documents. This is an instance of the Earth Mover's Distance (EMD).
Findings	The metric method leads to low error rates across all investigated data sets. WMD is also the among the slowest metrics to compute (a solution for speeding up the computations is presented in the article).
Challenges	The work points out that, (1) the interpretability of the method should be explored, and (2) the document structure should be considered using a distance function.
	Feng et al. [23]
	See Table 4
	Gomez and Moens [27]
Details	Classification inference is based on the reconstruction errors of each classification model for each class, i.e. measuring the difference between the set of reconstructed documents and the original one.
Findings	The proposed method creates a model that generalizes the classification problem well. Its performance depends on the number of principal components. The method performs better than the rest of the classifiers when a dataset has select properties.
Challenges	The work points out that, (1) other text classification tasks should be explored, and (2) the output prediction of the model should be combined with other classifiers to refine the final prediction.
	Lo and Ding [57]
Details	The article explores the background net [18, 56], and a set of different reasoning methods created on top of the net to resolve a document classification task.
Findings	The method produces impressive performance without demanding significant effort in preprocessing.

Challenges	The authors state that it is required to study how to obtain fuzzy association between terms based on granules of articles to achieve a more flexible and robust approach.
	Sainath et al. [70]
Details	The article compares three frameworks used to produce sparse coding solutions with different vocabulary sizes to generate a classification decision.
Findings	All training documents not only increase the size of the dictionary significantly, but also enforce a stronger need for sparseness on the coefficients. Sparse coding methods offer slight, but promising results over a Naive Bayes classifier.
Challenges	The work points out that, (1) feature selection techniques should be incorporated, and (2) comparison with other learning methods should be performed.
	Li and Vogel [49]
Details	The authors improve multi-class text classification using Error-Correcting Output Coding (ECOC) with sub-class partitions.
Findings	In ECOC, sub-class partition information of positive and negative classes is available, but ignored, even though it has a value for binary classification. No single algorithm can win on every dataset and situation.
Challenges	The work points out that, (1) more experiments on more datasets should be performed, (2) non-text applications should be considered, and (3) local search algorithms should be explored to improve the proposed strategy.
	Jin et al. [39]
Details	The authors create a new classification based on prototype learning, in which training data is represented as a set of points (prototypes) in a feature space.
Findings	The authors observed that the proposed method produces a larger average hypothesis margin than other prototype learning algorithms.
Challenges	The work points out that, the method can also be applied as a learning criterion to other classifier structures based on gradient descent, such as neural networks and quadratic discriminant functions.
	Xia et al. [93]
Details	The article explores the linear classification approach – a matrix of scores (the contribution table) is computed during the training process and a document is classified into the group with the largest score combination.
Findings	The method has lower time complexity, and does not need to know the semantic contribution of a term makes to a document in which it occurs.
Challenges	The work points out that, different feature weights should be considered.
	Larochelle and Bengio [47]
Details	The authors incorporate labels into the training process of Restricted Boltzmann Machines (RBMS), and propose two models: (1) Discriminative Restricted Boltzmann Machines (DRBMs), and (2) Hybrid Discriminative Restricted Boltzmann Machines (HDRBMs).
Findings	RBMs can and should be used as standalone non-linear classifiers. RBMs are effective at capturing the conditional statistical relationship between multiple tasks, or between the components in a complex target space.
Challenges	The work points out that, (1) more challenging settings, such as multi-task or structured output problems should be considered, (2) mean-field approximations should be applied, and (3) for large, but sparse input vectors, less computationally expensive learning should be introduced.

	Genkin et al. [25]
Details	A Laplace prior regularization term is used within a Bayesian logistic regression approach. An optimization method is also proposed.
Findings	The classification results depend on feature selection and configuration of the classification method. The authors found a strong correlation between the number of positive training examples and the number of features chosen.
Challenges	The work points out that, (1) other LASSO-based feature selection algorithms should be explored, and (2) scaling algorithms that estimate *regularization paths* to huge applications remain challenging.
	Qian et al. [65]
Details	The article employs the Associative Text Categorization (ATC) concept to produce a semantic-aware classifier, which includes understandable rules for text categorization.
Findings	The article confirms an observation from earlier research of [40]. A vertical rule-pruning method can greatly help reduce computational cost.
Challenges	The authors fail to highlight any challenges or open problems.
	Gliozzo et al. [26]
Details	The proposed algorithm utilizes a generalized similarity measure based on latent semantic spaces and a Gaussian Mixture algorithm to scale similarity scores into probabilities.
Findings	Competitive performance can be achieved only by using the category names as initial seeds.
Challenges	The work points out that, (1) the optimal procedures for collecting seed features should be investigated, (2) the contribution of additional seed performance should be explored, and (3) optimal combinations of intensional and extensional supervision should be investigated.
	Zhang et al. [103]
Details	The authors studied kernels on the multinomial manifold that enables Support Vector Machines (SVMs) to effectively exploit the intrinsic geometric structure of text data.
Findings	Negative Geodesic Distance (NGD) on the multinomial manifold is a conditionally positive definite (CPD) kernel, and leads to improvements in accuracy over kernels assuming Euclidean geometry. Linear kernel and TF-IDF with l_2 regularization achieve second result.
Challenges	The work points out that, (1) the NGD kernel should be extended to other manifolds (particularly for multimedia tasks), and (2) other kernel methods should be considered.
	Baoli et al. [7]
Details	The authors propose an improved and adaptive k-nearest neighbors strategy to resolve its problems.
Findings	The proposed methods are less sensitive to the parameter k, and can adequately classify documents belonging to smaller classes with larger values of k. The proposed strategy is adequate for cases in which estimating the parameter k via cross-validation is impossible, and the class distribution of the training set is skewed.

Challenges	The work points out that, (1) multi-label classification problems should be addressed, (2) the question of how to evaluate a dataset for text categorization should be addressed, and (3) a guideline on how to build a useful training collection for text categorization should be developed.
	Rennie [68]
Details	The work explores how the given validation method performs on a selection of regularization parameters of a classification method called Regularized Least Squares Classification (RLSC).
Findings	For RLSC, leave-one-out cross validation (LOOCV) consistently selects a regularization parameter that is too large.
Challenges	The work points out that, other text datasets should be considered.

1.7 Evaluation of learning algorithms and benchmarking methods

Table 7 lists four papers that collectively describe the evaluation problems of learning algorithms, and touch the issue of benchmarking methods.

Table 7. Evaluation and benchmarking articles.

Aspect of work	Reference/Description
	Wagh et al. [86]
Details	The authors' benchmark approaches range from simple Naive Bayes to complex BERT on six standard text classification datasets. They present an exhaustive comparison of different algorithms on various long document datasets.
Findings	Classifying long documents is a relatively simple task. Even basic algorithms can perform well compared to BERT-based approaches on most datasets. BERT-based models consistently perform well on all datasets, but their computational cost may be a concern. For shallower models, the authors recommend using a raw BiLSTM + Max architecture, which performs well across all datasets. A Glove + Attention bag of words model may suffice for more uncomplicated use cases. However, more sophisticated models are necessary for more challenging tasks such as the IMDB sentiment dataset.
Challenges	The work indicates that future work should focus on adequately incorporating conceptual labels into some NLP tasks.
	Suneera and Prakash [77]
Details	The authors evaluated the performance of various machine learning (ML) and deep learning algorithms for text classification. They chose six machine learning algorithms and three vectorization techniques, and five deep learning algorithms for the evaluation.
Findings	Results indicate that Logistic Regression outperforms other ML algorithms. A Bichannel Convolution Neural Network model gains exciting results compared to other deep learning models.
Challenges	The work points out that, Convolutional Neural Network (CNN) and Multilayer Perceptron (MLP) architectures may be notably improved by tuning their parameters and improving their basic architecture for various real-life applications.
	Bramesh and Anil Kumar [12]

Details	The performance of state-of-the-art classification methods that utilize vector space, in which terms weighted using weighting methods is compared.
Findings	The decision tree (C5.0) classifier performed well on all datasets examined.
Challenges	The work points out that, it should be extended to include other feature selection methods, and to utilize the k-fold validation procedure.
	Arras et al. [5]
Details	The authors demonstrate, based on two classification methods, that understanding of how and why a given classification method classifies can be achieved by tracing the classification decision back to individual words using layer-wise relevance propagation (LRP).
Findings	The proposed measure of a model's explanatory power depends only on the relevance of words. A CNN model produces better explanations than a BoW/SVM classifier, and incurs lower computational costs. The LRP decomposition method provides better explanations than gradient-based sensitivity analysis. A CNN can take advantage of the word similarity information encoded in the distributed word embeddings.
Challenges	The work points out that, the suitability of the model should be checked on other neural-based applications, or other types of classification problems, such as sentiment analysis.
	Mazyad et al. [59]
Details	The performance of state-of-the-art classification methods that utilize vector space, in which terms are weighted using feature weighting methods is compared.
Findings	The superiority of supervised term weighting methods over unsupervised methods remains unclear.
Challenges	The authors failed to highlight any challenges or open problems.
	Sun et al. [76]
Details	The performance of state-of-the-art strategies to address imbalanced text classification using SVMs is compared, and a survey of techniques proposed for imbalanced classification is presented.
Findings	SVMs learn the best decision surface in most test cases. For classification tasks involving high imbalance ratios, it is critical to find an appropriate threshold of SVMs.
Challenges	The work points out that, (1) better thresholding strategies should be developed, and (2) the learning objective function of the SVMs to consider the data imbalance in learning the decision surface should be improved.

REFERENCES

[1] Charu C. Aggarwal. 2016. *Outlier Analysis* (2nd ed.). Springer Publishing Company, Incorporated.

[2] Bassam Al-Salemi, Masri Ayob, and Shahrul Azman Mohd Noah. 2018. Feature ranking for enhancing boosting-based multi-label text categorization. *Expert Systems with Applications* 113 (2018), 531–543. https://doi.org/10.1016/j.eswa.2018.07.024

[3] Bassam Al-Salemi, Shahrul Azman Mohd Noah, and Mohd Juzaiddin Ab Aziz. 2016. RFBoost: An improved multi-label boosting algorithm and its application to text categorisation. *Knowledge-Based Systems* 103 (2016), 104–117. https://doi.org/10.1016/j.knosys.2016.03.029

[4] Ricardo Aler, José M. Valls, and Henrik Boström. 2020. Study of Hellinger Distance as a splitting metric for Random Forests in balanced and imbalanced classification datasets. *Expert Systems with Applications* 149 (7 2020), 113264. https://doi.org/10.1016/j.eswa.2020.113264

[5] Leila Arras, Franziska Horn, Grégoire Montavon, Klaus Robert Müller, and Wojciech Samek. 2017. "What is relevant in a text document?": An interpretable machine learning approach. *PLoS ONE* 12, 8 (2017), 1–23. https://doi.org/10.1371/journal.pone.0181142 arXiv:1612.07843

[6] Joseph Attieh and Joe Tekli. 2023. Supervised term-category feature weighting for improved text classification. *Knowledge-Based Systems* 261 (2 2023), 110215. https://doi.org/10.1016/j.knosys.2022.110215

[7] Li Baoli, Lu Qin, and Yu Shiwen. 2004. An adaptive k-nearest neighbor text categorization strategy. *ACM Transactions on Asian Language Information Processing* 3, 4 (2004), 215–226. https://doi.org/10.1145/1039621.1039623

[8] Fernando Benites and Elena Sapozhnikova. 2017. Improving scalability of ART neural networks. *Neurocomputing* 230 (2017), 219–229. https://doi.org/10.1016/j.neucom.2016.12.022

[9] Geir Thore Berge, Ole-Christoffer Granmo, Tor Oddbjorn Tveit, Morten Goodwin, Lei Jiao, and Bernt Viggo Matheussen. 2019. Using the Tsetlin Machine to Learn Human-Interpretable Rules for High-Accuracy Text Categorization With Medical Applications. *IEEE Access* 7 (2019), 115134–115146. https://doi.org/10.1109/access.2019.2935416 arXiv:1809.04547

[10] Marcin Białas, Marcin Michał Mirończuk, and Jacek Mańdziuk. 2020. Biologically Plausible Learning of Text Representation with Spiking Neural Networks. *Parallel Problem Solving from Nature PPSN XVI*, 433–447. https://doi.org/10.1007/978-3-030-58112-1_30

[11] Charles Bouveyron, Gilles Celeux, T. Brendan Murphy, and Adrian E. Raftery. 2019. *Model-Based Clustering and Classification for Data Science: With Applications in R.* Cambridge University Press. https://doi.org/10.1017/9781108644181

[12] S. M. Bramesh and K. M. Anil Kumar. 2019. Empirical Study to Evaluate the Performance of Classification Algorithms on Public Datasets. In *Lecture Notes in Electrical Engineering*. Vol. 545. Springer, 447–455. https://doi.org/10.1007/978-981-13-5802-9_41

[13] Austin J. Brockmeier, Tingting Mu, Sophia Ananiadou, and John Y. Goulermas. 2018. Self-Tuned Descriptive Document Clustering Using a Predictive Network. *IEEE Transactions on Knowledge and Data Engineering* 30, 10 (2018), 1929–1942. https://doi.org/10.1109/TKDE.2017.2781721

[14] Deng Cai and Xiaofei He. 2012. Manifold adaptive experimental design for text categorization. *IEEE Transactions on Knowledge and Data Engineering* 24, 4 (2012), 707–719. https://doi.org/10.1109/TKDE.2011.104

[15] C. L. Philip Chen and Shuang Feng. 2020. Generative and Discriminative Fuzzy Restricted Boltzmann Machine Learning for Text and Image Classification. *IEEE Transactions on Cybernetics* 50 (5 2020), 2237–2248. Issue 5. https://doi.org/10.1109/TCYB.2018.2869902

[16] Gang Chen and Sargur N. Srihari. 2020. Revisiting hierarchy: Deep learning with orthogonally constrained prior for classification. *Pattern Recognition Letters* 140 (12 2020), 214–221. https://doi.org/10.1016/j.patrec.2020.10.006

[17] Kewen Chen, Zuping Zhang, Jun Long, and Hao Zhang. 2016. Turning from TF-IDF to TF-IGM for term weighting in text classification. *Expert Systems with Applications* 66 (2016), 1339–1351. https://doi.org/10.1016/j.eswa.2016.09.009

[18] Yuan Chen, Li Ya Ding, Sio Long Lo, and Wenjun Sun. 2011. Background net for personalized keywords in article selection. *ICIC Express Letters* 5, 4 B (2011), 1339–1346.

[19] Yu Chen and Mohammed J. Zaki. 2017. KATE: K-competitive autoencoder for text. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Vol. Part F1296. ACM, 85–94. https://doi.org/10.1145/3097983.3098017 arXiv:1705.02033

[20] Billy Chiu, Sunil Kumar Sahu, Neha Sengupta, Derek Thomas, and Mohammady Mahdy. 2020. Attending to Inter-sentential Features in Neural Text Classification. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1685–1688. https://doi.org/10.1145/3397271.3401203

[21] Yong Dai, Linjun Shou, Ming Gong, Xiaolin Xia, Zhao Kang, Zenglin Xu, and Daxin Jiang. 2022. Graph Fusion Network for Text Classification. *Knowledge-Based Systems* 236 (1 2022), 107659. https://doi.org/10.1016/j.knosys.2021.107659

[22] Kaize Ding, Jianling Wang, Jundong Li, Dingcheng Li, and Huan Liu. 2020. Be More with Less: Hypergraph Attention Networks for Inductive Text Classification. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4927–4936. https://doi.org/10.18653/v1/2020.emnlp-main.399

[23] Guozhong Feng, Jianhua Guo, Bing Yi Jing, and Tieli Sun. 2015. Feature subset selection using naive Bayes for text classification. *Pattern Recognition Letters* 65 (2015), 109–115. https://doi.org/10.1016/j.patrec.2015.07.028

[24] Xiaoyue Feng, Yanchun Liang, Xiaohu Shi, Dong Xu, Xu Wang, and Renchu Guan. 2017. Overfitting reduction of text classification based on AdaBELM. *Entropy* 19, 7 (2017), 330. https://doi.org/10.3390/e19070330

[25] Alexander Genkin, David D. Lewis, and David Madigan. 2007. Large-scale bayesian logistic regression for text categorization. *Technometrics* 49, 3 (2007), 291–304. https://doi.org/10.1198/004017007000000245

[26] Alfio Gliozzo, Carlo Strapparava, and Ido Dagan. 2005. Investigating unsupervised learning for text categorization bootstrapping. In *HLT/EMNLP 2005 - Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. The Association for Computational Linguistics, 129–136. https://doi.org/10.3115/1220575.1220592

[27] Juan Carlos Gomez and Marie Francine Moens. 2014. Minimizer of the Reconstruction Error for multi-class document categorization. *Expert Systems with Applications* 41, 3 (2014), 861–868. https://doi.org/10.1016/j.eswa.2013.08.016

[28] Emanuele Guidotti and Alfio Ferrara. 2022. Text Classification with Born's Rule, Alice H Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). *Advances in Neural Information Processing Systems.* https://openreview.net/forum?id=sNcn-E3uPHA

[29] Shun Guo and Nianmin Yao. 2020. Generating word and document matrix representations for document classification. *Neural Computing and Applications* 32 (7 2020), 10087–10108. Issue 14. https://doi.org/10.1007/s00521-019-04541-x

[30] Shun Guo and Nianmin Yao. 2021. Document Vector Extension for Documents Classification. *IEEE Transactions on Knowledge and Data Engineering* 33 (8 2021), 3062–3074. Issue 8. https://doi.org/10.1109/TKDE.2019.2961343

[31] Vivek Gupta, Ankit Kumar Saw, Pegah Nokhiz, Harshit Gupta, and Partha Pratim Talukdar. 2019. Improving Document Classification with Multi-Sense Embeddings. *European Conference on Artificial Intelligence.*

[32] Abdelaali Hassaine, Zeineb Safi, Jameela Otaibi, and Ali Jaoua. 2017. Text categorization using weighted hyper rectangular keyword extraction. In *Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA*, Vol. 2017-Octob. {IEEE} Computer Society, 959–965. https://doi.org/10.1109/AICCSA.2017.102

[33] Geoffrey E. Hinton, Simon Osindero, and Yee Whye Teh. 2006. A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation* 18, 7 (2006), 1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527

[34] Junying Hu, Jiangshe Zhang, Nannan Ji, and Chunxia Zhang. 2017. A new regularized restricted Boltzmann machine based on class preserving. *Knowledge-Based Systems* 123 (2017), 1–12. https://doi.org/10.1016/j.knosys.2017.02.012

[35] Kashif Javed and Haroon A. Babri. 2017. Feature selection based on a normalized difference measure for text classification. *Information Processing and Management* 53, 2 (2017), 473–489. https://doi.org/10.1016/j.ipm.2016.12.004

[36] Longjia Jia and Bangzuo Zhang. 2022. A new document representation based on global policy for supervised term weighting schemes in text categorization. *Mathematical Biosciences and Engineering* 19 (2022), 5223–5240. Issue 5. https://doi.org/10.3934/mbe.2022245

[37] Haiyun Jiang, Deqing Yang, Yanghua Xiao, and Wei Wang. 2020. Understanding a bag of words by conceptual labeling with prior weights. *World Wide Web* 23 (7 2020), 2429–2447. Issue 4. https://doi.org/10.1007/s11280-020-00806-x

[38] Peng Jin, Zhang Yue, Xingyuan Chen, and Yunqing Xia. 2016. Bag-of-embeddings for text classification. In *IJCAI International Joint Conference on Artificial Intelligence*, Subbarao Kambhampati (Ed.), Vol. 2016-Janua. {IJCAI/AAAI} Press, 2824–2830. http://www.ijcai.org/Abstract/16/401

[39] Xiao Bo Jin, Cheng Lin Liu, and Xinwen Hou. 2010. Regularized margin-based conditional log-likelihood loss for prototype learning. *Pattern Recognition* 43, 7 (2010), 2428–2438. https://doi.org/10.1016/j.patcog.2010.01.013

[40] Thorsten Joachims. 1997. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML 1997), Nashville, Tennessee, USA, July 8-12, 1997*, Douglas H. Fisher (Ed.). Morgan Kaufmann, 143–151.

[41] Rie Johnson and Tong Zhang. 2016. Supervised and Semi-Supervised Text Categorization Using LSTM for Region Embeddings. *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, 526–534.

[42] Santosh Kesiraju, Lukáš Burget, Igor Szoke, and Jan Černocký. 2016. Learning document representations using subspace multinomial model. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Nelson Morgan (Ed.), Vol. 08-12-Sept. ISCA, 700–704. https://doi.org/10.21437/Interspeech.2016-1634

[43] Santosh Kesiraju, Oldrich Plchot, Lukas Burget, and Suryakanth V. Gangashetty. 2020. Learning Document Embeddings Along With Their Uncertainties. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), 2319–2332. https://doi.org/10.1109/TASLP.2020.3012062

[44] Snehal Ishwar Khandve, Vedangi Kishor Wagh, Apurva Dinesh Wani, Isha Mandar Joshi, and Raviraj Bhuminand Joshi. 2022. Hierarchical Neural Network Approaches for Long Document Classification. *2022 14th International Conference on Machine Learning and Computing (ICMLC)*, 115–119. https://doi.org/10.1145/3529836.3529935

[45] Donghwa Kim, Deokseong Seo, Suhyoun Cho, and Pilsung Kang. 2019. Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec. *Information Sciences* 477 (2019), 15–29. https://doi.org/10.1016/j.ins.2018.10.006

[46] Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *32nd International Conference on Machine Learning, ICML 2015 ({JMLR} Workshop and Conference Proceedings, Vol. 2)*, Francis R Bach and David M Blei (Eds.). JMLR.org, 957–966. http://proceedings.mlr.press/v37/kusnerb15.html

[47] Hugo Larochelle and Yoshua Bengio. 2008. Classification using discriminative restricted boltzmann machines. In *Proceedings of the 25th International Conference on Machine Learning ({ACM} International Conference Proceeding Series, Vol. 307)*, William W Cohen, Andrew McCallum, and Sam T Roweis (Eds.). ACM, 536–543. https://doi.org/10.1145/1390156.1390224

[48] Baoli Li. 2016. Importance weighted feature selection strategy for text classification. In *Proceedings of the 2016 International Conference on Asian Language Processing, IALP 2016*, Minghui Dong, Yuen-Hsien Tseng, Yanfeng Lu, Liang-Chih Yu, Lung-Hao Lee, Chung-Hsien Wu, and Haizhou Li (Eds.). IEEE, 344–347. https://doi.org/10.1109/IALP.2016.7876002

[49] Baoli Li and Carl Vogel. 2010. Improving multiclass text classification with error-correcting output coding and sub-class partitions. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Lecture Notes in Computer Science, Vol. 6085 LNAI)*, Atefeh Farzindar and Vlado Keselj (Eds.). Springer, 4–15. https://doi.org/10.1007/978-3-642-13059-5_4

[50] Baoli Li, Qiuling Yan, Zhenqiang Xu, and Guicai Wang. 2015. Weighted Document Frequency for feature selection in text classification. In *Proceedings of 2015 International Conference on Asian Language Processing, IALP 2015*. IEEE, 132–135. https://doi.org/10.1109/IALP.2015.7451549

[51] Jianqiang Li, Jing Li, Xianghua Fu, M. A. Masud, and Joshua Zhexue Huang. 2016. Learning distributed word representation with multi-contextual mixed embedding. *Knowledge-Based Systems* 106 (2016), 220–230. https://doi.org/10.1016/j.knosys.2016.05.045

[52] Pengfei Li, Kezhi Mao, Yuecong Xu, Qi Li, and Jiaheng Zhang. 2020. Bag-of-Concepts representation for document classification based on automatic knowledge acquisition from probabilistic knowledge base. *Knowledge-Based Systems* 193 (2020), 105436. https://doi.org/10.1016/j.knosys.2019.105436

[53] Zhixing Li, Zhongyang Xiong, Yufang Zhang, Chunyong Liu, and Kuan Li. 2011. Fast text categorization using concise semantic analysis. *Pattern Recognition Letters* 32, 3 (2011), 441–448. https://doi.org/10.1016/j.patrec.2010.11.001

[54] Yuxiao Lin, Yuxian Meng, Xiaofei Sun, Qinghong Han, Kun Kuang, Jiwei Li, and Fei Wu. 2021. BertGCN: Transductive Text Classification by Combining GNN and BERT. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 1456–1462. https://doi.org/10.18653/v1/2021.findings-acl.126

[55] Xien Liu, Xinxin You, Xiao Zhang, Ji Wu, and Ping Lv. 2020. Tensor Graph Convolutional Networks for Text Classification. *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (4 2020), 8409–8416. Issue 05. https://doi.org/10.1609/aaai.v34i05.6359

[56] Sio-Long Lo, Liya Ding, and Yuan Chen. 2011. Incremental learning on background net to capture changing personal reading preference. In *International Conference on Machine Learning and Cybernetics, ICMLC 2011, Guilin, China, July 10-13, 2011, Proceedings*. IEEE, 1503–1508. https://doi.org/10.1109/ICMLC.2011.6016972

[57] Sio Long Lo and Liya Ding. 2012. Probabilistic reasoning on background net: An application to text categorization. In *Proceedings - International Conference on Machine Learning and Cybernetics*, Vol. 2. IEEE, 688–694. https://doi.org/10.1109/ICMLC.2012.6359008

[58] Qiming Luo, Enhong Chen, and Hui Xiong. 2011. A semantic term weighting scheme for text categorization. *Expert Systems with Applications* 38, 10 (2011), 12708–12716. https://doi.org/10.1016/j.eswa.2011.04.058

[59] Ahmad Mazyad, Fabien Teytaud, and Cyril Fonlupt. 2017. A comparative study on term weighting schemes for text classification. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Lecture Notes in Computer Science, Vol. 10710 LNCS)*, Giuseppe Nicosia, Panos M Pardalos, Giovanni Giuffrida, and Renato Umeton (Eds.). Springer, 100–108. https://doi.org/10.1007/978-3-319-72926-8_9

[60] Dinesh Nagumothu, Peter W. Eklund, Bahadorreza Ofoghi, and Mohamed Reda Bouadjenek. 2021. Linked Data Triples Enhance Document Relevance Classification. *Applied Sciences* 11 (7 2021), 6636. Issue 14. https://doi.org/10.3390/app11146636

[61] Guansong Pang, Huidong Jin, and Shengyi Jiang. 2015. CenKNN: a scalable and effective text classifier. *Data Mining and Knowledge Discovery* 29, 3 (2015), 593–625. https://doi.org/10.1007/s10618-014-0358-x

[62] Raghavendra Pappagari, Jesus Villalba, and Najim Dehak. 2018. Joint Verification-Identification in end-to-end Multi-Scale CNN Framework for Topic Identification. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Vol. 2018-April. IEEE, 6199–6203. https://doi.org/10.1109/ICASSP.2018.8461673

[63] Miha Pavlinek and Vili Podgorelec. 2017. Text classification method based on self-training and LDA topic models. *Expert Systems with Applications* 80 (2017), 83–93. https://doi.org/10.1016/j.eswa.2017.03.020

[64] Sunil Kumar Prabhakar, Harikumar Rajaguru, Kwangsub So, and Dong-Ok Won. 2022. A Framework for Text Classification Using Evolutionary Contiguous Convolutional Neural Network and Swarm Based Deep Neural Network. *Frontiers in Computational Neuroscience* 16 (6 2022). https://doi.org/10.3389/fncom.2022.900885

[65] Tieyun Qian, Hui Xiong, Yuanzhen Wang, and Enhong Chen. 2007. On the strength of hyperclique patterns for text categorization. *Information Sciences* 177, 19 (2007), 4040–4058. https://doi.org/10.1016/j.ins.2007.04.005

[66] Rahul Ragesh, Sundararajan Sellamanickam, Arun Iyer, Ramakrishna Bairi, and Vijay Lingam. 2021. HeteGCN: Heterogeneous Graph Convolutional Networks for Text Classification. *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 860–868. https://doi.org/10.1145/3437963.3441746

[67] Abdur Rehman, Kashif Javed, Haroon A. Babri, and Mehreen Saeed. 2015. Relative discrimination criterion - A novel feature ranking method for text data. *Expert Systems with Applications* 42, 7 (2015), 3670–3681. https:

//doi.org/10.1016/j.eswa.2014.12.013

[68] Jason D M Rennie. 2003. *On the value of leave-one-out cross-validation bounds.* Technical Report. https://pdfs.semanticscholar.org/4f8e/c2f5b9d6dc1bbccf6d8c03730c3cec4e35f7.pdf

[69] Thiago Fredes Rodrigues and Paulo Martins Engel. 2014. Probabilistic clustering and classification for textual data: An online and incremental approach. In *Proceedings - 2014 Brazilian Conference on Intelligent Systems, BRACIS 2014*. {IEEE} Computer Society, 288–293. https://doi.org/10.1109/BRACIS.2014.59

[70] Tara N. Sainath, Sameer Maskey, Dimitri Kanevsky, Bhuvana Ramabhadran, David Nahamoo, and Julia Hirschberg. 2010. Sparse representations for text categorization. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, Takao Kobayashi, Keikichi Hirose, and Satoshi Nakamura (Eds.). ISCA, 2266–2269. https://doi.org/10/i10

[71] Ruslan Salakhutdinov and Geoffrey Hinton. 2009. Replicated softmax: An undirected topic model. In *Advances in Neural Information Processing Systems 22 - Proceedings of the 2009 Conference*, Yoshua Bengio, Dale Schuurmans, John D Lafferty, Christopher K I Williams, and Aron Culotta (Eds.). Curran Associates, Inc., 1607–1614. http://papers.nips.cc/paper/3856-replicated-softmax-an-undirected-topic-model

[72] Neeraj Sharma, A. D. Dileep, and Veena Thenkanidiyoor. 2017. Text classification using hierarchical sparse representation classifiers. In *Proceedings - 16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017*, Vol. 2017-Decem. IEEE, 1015–1019. https://doi.org/10.1109/ICMLA.2017.00-18

[73] Neeraj Sharma, Anshu Sharma, Veena Thenkanidiyoor, and A. D. Dileep. 2016. Text classification using combined sparse representation classifiers and support vector machines. In *2016 4th International Symposium on Computational and Business Intelligence, ISCBI 2016*. IEEE, 181–185. https://doi.org/10.1109/ISCBI.2016.7743280

[74] Farhan Shehzad, Abdur Rehman, Kashif Javed, Khalid A. Alnowibet, Haroon A. Babri, and Hafiz Tayyab Rauf. 2022. Binned Term Count: An Alternative to Term Frequency for Text Categorization. *Mathematics* 10 (11 2022), 4124. Issue 21. https://doi.org/10.3390/math10214124

[75] Yuan-Yuan Shen, Yan-Ming Zhang, Xu-Yao Zhang, and Cheng-Lin Liu. 2020. Online semi-supervised learning with learning vector quantization. *Neurocomputing* 399 (7 2020), 467–478. https://doi.org/10.1016/j.neucom.2020.03.025

[76] Aixin Sun, Ee Peng Lim, and Ying Liu. 2009. On strategies for imbalanced text classification using SVM: A comparative study. *Decision Support Systems* 48, 1 (2009), 191–201. https://doi.org/10.1016/j.dss.2009.07.011

[77] C M Suneera and Jay Prakash. 2020. Performance Analysis of Machine Learning and Deep Learning Models for Text Classification. *2020 IEEE 17th India Council International Conference (INDICON)*, 1–6. https://doi.org/10.1109/INDICON49873.2020.9342208

[78] Bo Tang, Haibo He, Paul M. Baggenstoss, and Steven Kay. 2016. A Bayesian Classification Approach Using Class-Specific Features for Text Categorization. *IEEE Transactions on Knowledge and Data Engineering* 28, 6 (2016), 1602–1606. https://doi.org/10.1109/TKDE.2016.2522427

[79] Bo Tang, Steven Kay, and Haibo He. 2016. Toward Optimal Feature Selection in Naive Bayes for Text Categorization. *IEEE Transactions on Knowledge and Data Engineering* 28, 9 (2016), 2508–2521. https://doi.org/10.1109/TKDE.2016.2563436 arXiv:1602.02850

[80] Zhong Tang, Wenqiang Li, and Yan Li. 2020. An improved term weighting scheme for text classification. *Concurrency Computation* 32, 9 (2020). https://doi.org/10.1002/cpe.5604

[81] Zhong Tang, Wenqiang Li, and Yan Li. 2022. An improved supervised term weighting scheme for text representation and classification. *Expert Systems with Applications* 189 (3 2022), 115985. https://doi.org/10.1016/j.eswa.2021.115985

[82] Roman Tesar, Massimo Poesio, Vaclav Strnad, and Karel Jezek. 2006. Extending the single words-based document model: A comparison of bigrams and 2-itemsets. In *Proceedings of the 2006 ACM Symposium on Document Engineering, DocEng 2006*, Dick C A Bulterman and David F Brailsford (Eds.), Vol. 2006. ACM, 138–146. https://doi.org/10.1145/1166160.1166197

[83] Prayag Tiwari and Massimo Melucci. 2019. Towards a Quantum-Inspired Binary Classifier. *IEEE Access* 7 (2019), 42354–42372. https://doi.org/10.1109/ACCESS.2019.2904624

[84] Narendra Babu Unnam and P. Krishna Reddy. 2020. A document representation framework with interpretable features using pre-trained word embeddings. *International Journal of Data Science and Analytics* 10 (6 2020), 49–64. Issue 1. https://doi.org/10.1007/s41060-019-00200-5

[85] P. Unnikrishnan, V. K. Govindan, and S. D. Madhu Kumar. 2019. Enhanced sparse representation classifier for text classification. *Expert Systems with Applications* 129 (2019), 260–272. https://doi.org/10.1016/j.eswa.2019.04.003

[86] Vedangi Wagh, Snehal Khandve, Isha Joshi, Apurva Wani, Geetanjali Kale, and Raviraj Joshi. 2021. Comparative Study of Long Document Classification. *TENCON 2021 - 2021 IEEE Region 10 Conference (TENCON)*, 732–737. https://doi.org/10.1109/TENCON54134.2021.9707465

[87] Fen Wang, Xiaoxuan Li, Xiaotao Huang, and Ling Kang. 2016. Improved document feature selection with categorical parameter for text classification. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Lecture Notes in Computer Science, Vol. 10026 LNCS)*, Selma Boumerdassi,

Éric Renault, and Samia Bouzefrane (Eds.). Springer, 86–98. https://doi.org/10.1007/978-3-319-50463-6_8

[88] Shuaihui Wang, Yu Pan, Jin Zhang, Xingyu Zhou, Zhen Cui, Guyu Hu, and Zhisong Pan. 2021. Robust and label efficient bi-filtering graph convolutional networks for node classification. *Knowledge-Based Systems* 224 (7 2021), 106891. https://doi.org/10.1016/j.knosys.2021.106891

[89] Tao Wang, Yi Cai, Ho fung Leung, Raymond Y. K. Lau, Haoran Xie, and Qing Li. 2021. On entropy-based term weighting schemes for text categorization. *Knowledge and Information Systems* 63 (9 2021), 2313–2346. Issue 9. https://doi.org/10.1007/s10115-021-01581-5

[90] Yizhao Wang, Chenxi Wang, Jieyu Zhan, Wenjun Ma, and Yuncheng Jiang. 2023. Text FCG: Fusing Contextual Information via Graph Learning for text classification. *Expert Systems with Applications* 219 (6 2023), 119658. https://doi.org/10.1016/j.eswa.2023.119658

[91] Zhengjue Wang, Chaojie Wang, Hao Zhang, Zhibin Duan, Mingyuan Zhou, and Bo Chen. 2020. Learning Dynamic Hierarchical Topic Graph with Graph Convolutional Network for Document Classification, Silvia Chiappa and Roberto Calandra (Eds.). *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics* 108, 3959–3969. https://proceedings.mlr.press/v108/wang20l.html

[92] Xinde Wei, Hai Huang, Longxuan Ma, Ze Yang, and Liutong Xu. 2020. Recurrent Graph Neural Networks for Text Classification. *2020 IEEE 11th International Conference on Software Engineering and Service Science (ICSESS)*, 91–97. https://doi.org/10.1109/ICSESS49938.2020.9237709

[93] Feng Xia, Tian Jicun, and Liu Zhihui. 2009. A text categorization method based on local document frequency. In *6th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2009 (FSKD'09, Vol. 7)*. IEEE Press, 468–471. https://doi.org/10.1109/FSKD.2009.291

[94] Qianqian Xie, Jimin Huang, Pan Du, Min Peng, and Jian-Yun Nie. 2021. Graph Topic Neural Network for Document Representation. *Proceedings of the Web Conference 2021*, 3055–3065. https://doi.org/10.1145/3442381.3450045

[95] Qianqian Xie, Jimin Huang, Pan Du, Min Peng, and Jian-Yun Nie. 2021. Inductive Topic Variational Graph Auto-Encoder for Text Classification. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4218–4227. https://doi.org/10.18653/v1/2021.naacl-main.333

[96] Jun Yan, Ning Liu, Qiang Yang, Weiguo Fan, and Zheng Chen. 2008. TOFA: Trace oriented feature analysis in text categorization. In *Proceedings - IEEE International Conference on Data Mining, ICDM*. {IEEE} Computer Society, 668–677. https://doi.org/10.1109/ICDM.2008.67

[97] Peng Yan, Linjing Li, Miaotianzi Jin, and Daniel Zeng. 2021. Quantum probability-inspired graph neural network for document representation and classification. *Neurocomputing* 445 (7 2021), 276–286. https://doi.org/10.1016/j.neucom.2021.02.060

[98] Fei Yang, Huyin Zhang, Shiming Tao, and Sheng Hao. 2022. Graph representation learning via simple jumping knowledge networks. *Applied Intelligence* 52 (8 2022), 11324–11342. Issue 10. https://doi.org/10.1007/s10489-021-02889-z

[99] Gao Yang, Wang Wenbo, Liu Qian, Huang Heyan, and Yuefeng Li. 2018. Extending Embedding Representation by Incorporating Latent Relations. *IEEE Access* 6 (2018), 52682–52690. https://doi.org/10.1109/ACCESS.2018.2866531

[100] Liang Yang, Fan Wu, Junhua Gu, Chuan Wang, Xiaochun Cao, Di Jin, and Yuanfang Guo. 2020. Graph Attention Topic Modeling Network. *Proceedings of The Web Conference 2020*, 144–154. https://doi.org/10.1145/3366423.3380102

[101] Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph Convolutional Networks for Text Classification. *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (7 2019), 7370–7377. Issue 01. https://doi.org/10.1609/aaai.v33i01.33017370

[102] Cheng Zhang and Hayato Yamana. 2021. Improving Text Classification Using Knowledge in Labels. *2021 IEEE 6th International Conference on Big Data Analytics (ICBDA)*, 193–197. https://doi.org/10.1109/ICBDA51983.2021.9403092

[103] Dell Zhang, Xi Chen, and Wee Sun Lee. 2005. Text classification with kernels on the multinomial manifold. In *SIGIR 2005 - Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Ricardo A Baeza-Yates, Nivio Ziviani, Gary Marchionini, Alistair Moffat, and John Tait (Eds.). ACM, 266–273. https://doi.org/10.1145/1076034.1076081

[104] Suncong Zheng, Hongyun Bao, Jiaming Xu, Yuexing Hao, Zhenyu Qi, and Hongwei Hao. 2016. A Bidirectional Hierarchical Skip-Gram model for text topic embedding. In *Proceedings of the International Joint Conference on Neural Networks*, Vol. 2016-Octob. IEEE, 855–862. https://doi.org/10.1109/IJCNN.2016.7727289

[105] Yuxiang Zhou, Lejian Liao, Yang Gao, and Heyan Huang. 2021. Extracting salient features from convolutional discriminative filters. *Information Sciences* 558 (5 2021), 265–279. https://doi.org/10.1016/j.ins.2020.12.084

[106] Yuxiang Zhou, Lejian Liao, Yang Gao, Heyan Huang, and Xiaochi Wei. 2020. A Discriminative Convolutional Neural Network with Context-aware Attention. *ACM Transactions on Intelligent Systems and Technology* 11 (10 2020), 1–21. Issue 5. https://doi.org/10.1145/3397464

[107] Hao Zhu and Piotr Koniusz. 2021. Simple Spectral Graph Convolution. *International Conference on Learning Representations*. https://openreview.net/forum?id=CYO5T-YjWZV

[108] Wei Zong, Feng Wu, Lap Keung Chu, and Domenic Sculli. 2015. A discriminative and semantic feature selection method for text categorization. *International Journal of Production Economics* 165 (2015), 215–222. https://doi.org/10.1016/j.ijpe.2014.12.035

Received xx xx xxxx; revised xx xx xxxx; accepted xx xx xxxx