Artificial Intelligence Review (2023) 56:9401–9469
https://doi.org/10.1007/s10462-023-10393-8

A review of semi-supervised learning for text classification

José Marcio Duarte[1] · **Lilian Berton**[1]

Published online: 31 January 2023

Abstract
A huge amount of data is generated daily leading to big data challenges. One of them is related to text mining, especially text classification. To perform this task we usually need a large set of labeled data that can be expensive, time-consuming, or difficult to be obtained. Considering this scenario semi-supervised learning (SSL), the branch of machine learning concerned with using labeled and unlabeled data has expanded in volume and scope. Since no recent survey exists to overview how SSL has been used in text classification, we aim to fill this gap and present an up-to-date review of SSL for text classification. We retrieve 1794 works from the last 5 years from IEEE Xplore, ACM Digital Library, Science Direct, and Springer. Then, 157 articles were selected to be included in this review. We present the application domain, datasets, and languages employed in the works. The text representations and machine learning algorithms. We also summarize and organize the works following a recent taxonomy of SSL. We analyze the percentage of labeled data used, the evaluation metrics, and obtained results. Lastly, we present some limitations and future trends in the area. We aim to provide researchers and practitioners with an outline of the area as well as useful information for their current research.

Keywords Natural language processing · Text classification · Machine learning · Semi-supervised learning

1 Introduction

The number of internet users globally will grow to 5.3 billion by 2023 according to estimates of Statista (2022). The digital world promotes myriad downloads and data uploads, especially in text format. A significant amount of text data has been produced, it is time-consuming, expensive, and difficult to perform a manual curation of this content. Being desired algorithms that automatically classify documents and assist text mining tasks (Hassani et al. 2020).

✉ Lilian Berton
lberton@unifesp.br

José Marcio Duarte
josmarcioduarte@gmail.com

[1] Science and Technology Department, Federal University of São Paulo, Cesare Mansueto Giulio Lattes Ave, 1201, São José dos Campos, SP 12247-014, Brazil

Since texts are a rich source of information, techniques that automatically analyze and structure text cost-effectively are of great interest for academic and business applications. Text classification aims to assign a predefined category to a text. It is one of the principal tasks in Natural Language Processing (NLP) with several applications. Text classification is usually divided into supervised, unsupervised, and semi-supervised approaches (Thangaraj and Sivakami 2018). Supervised is the most expensive since depends on labeled data, the most common algorithms explored are SVM, decision tree, KNN, and neural networks. Unsupervised learning occurs when labeled data is not accessible, and the performance is not always good. The most common algorithms explored are K-Means, hierarchical clustering, and fuzzy c-means. Semi-supervised are used when there are few labeled data and a lot of unlabeled data. Common algorithms explored are co-training, self-training, transductive SVM, and graph-based methods.

Training data is a bottleneck in text classification and a great challenge is the labeling process, which involves a human annotator, who interprets and categorizes the content. This is time-consuming and expensive. So, machine learning techniques such as semi-supervised learning (SSL) that consider few labeled data and allow it to scale to any application can enable real-time analysis. This way, semi-supervised approaches become a hot research topic that uses the few labeled data and then classifies unlabeled documents (Zhou 2021; Van Engelen and Hoos 2020). NLP and SSL techniques have been combined and employed in different domains such as sentiment analysis (Silva et al. 2016; Han et al. 2020; Lee et al. 2019), word sense disambiguation (Duarte et al. 2021; Li et al. 2019), fake news detection (Benamira et al. 2019) and text classification (Linmei et al. 2019; Alam et al. 2018), reaching interesting results.

The idea of combining labeled and unlabeled data has been investigated for a long time, it starts in the statistic area when some authors proposed building classifiers with likelihood maximization by testing all possible class assignments (Hartley and Rao 1968; Day 1969). Since then, different approaches have been proposed for SSL. In Van Engelen and Hoos (2020) a taxonomy was proposed dividing the techniques into wrapper methods, unsupervised preprocessing, intrinsically semi-supervised, and graph-based. In the last years, some surveys presented text classification algorithms (Kowsari et al. 2019; Kadhim 2019), the main deep learning approaches used in text classification (Minaee et al. 2021), feature selection techniques (Deng et al. 2019), however, as far as we know no review has focused in SSL techniques for text classification, which are our focus.

One of the problems investigated in the SSL is the classifier degradation performance concerning the unlabeled data quantity added to a fixed set of labeled data. Nigam et al. (2000) used expectation-maximization (EM) combined with a generative classifier to investigate unlabeled and labeled samples in text classification. Cozman and Cohen (2002) analyzed the maximum-likelihood estimator and generative classifier focusing on modeling errors to evaluate the effect of unlabeled samples. More recently, Banitalebi-Dehkordi et al. (2022) showed that the unlabeled data from unconstrained distributions can generate a drop in the accuracy of SSL methods.

Text classification is the basis of many applications already mentioned, such as sentiment analysis, spam and fraud detection, word sense disambiguation, and so on, becoming a big issue in the field of artificial intelligence. This paper aims to retrieve and analyze the main approaches for text classification, especially, employing SSL, and presents their strengths and weaknesses. This study is very important for the computer science area, to help researchers and professionals to know the current research trends, develop customized

models, and support project development and knowledge discovery. The information condensed here can help to optimize resources and maximize accuracy.

This work is an endeavor to retrieve and contextualize the main approaches of SSL for text classification, as well as its recent advances. We access the publications from the last 5 years in four digital libraries (ACM, IEEE Xplore, Science Direct, and Springer). Initially, we selected 1794 articles, after applying exclusion criteria, 157 articles were chosen to be included in this review. The main contributions of this work are: (i) identify the main idioms, domains and tasks explored; (ii) retrieve the datasets used; (iii) identify the primary text representation used; (iv) detect the main algorithms used; (v) organize the works into the SSL techniques; (vi) find the percentage of labeled data and the results achieved by the SSL approaches into the datasets; (vii) present the strengths, limitations, and current research trends in SSL text classification.

Section 2 presents the methodology employed to retrieve and select the articles to be included in this review. Section 3 shows the results in graphic format to facilitate the view and interpretation, besides a brief discussion of the results. Section 4 presents the works divided into the main SSL approaches. Section 5 shows a comparative analysis of the results obtained by the works in the main tasks and datasets. Section 6 presents the benefits and limitations of the techniques. Section 7 presents the future opportunities in the area and Section 8 concludes the paper.

2 Methodology

This section presents the methodology employed to perform the review and retrieve the works. Section 2.1 presents the research question that guided the work. Section 2.2 presents the sources and search terms used for the research. Section 2.3 presents the selection process and exclusion criteria. Section 2.4 presents the main information we extract from the articles.

2.1 Research question

Principal question: Which are the approaches in the semi-supervised text classification that achieved relevant results in recent years?

To answer the main question we constructed the knowledge-based considering the used semi-supervised approaches, text representations, text classification tasks, machine learning algorithms, languages, domains, and datasets.

2.2 Sources and search terms

First, we performed the search at the end of March 2021 on four digital libraries, taking into account publications from the previous 5 years: ACM Digital Library,[1] IEEE Xplore,[2] Science Direct[3] and Springer.[4] **Title**, **abstract**, and **keywords** are the fields that we used

[1] ACM Digital Library: http://portal.acm.org/.
[2] IEEE Xplore: http://ieeexplore.ieee.org/.
[3] Science Direct: http://www.sciencedirect.com/.
[4] https://link.springer.com/.

to elaborate the search expression to select the articles on ACM and Science Direct. We used the field **All Metadata** on IEEE Xplore. However, in the Springer library, the number of articles returned was much bigger than the other libraries considering that it does not separate the articles by fields, instead, all the text is analyzed. The search expression used on the libraries were: *(text classification) AND (semi-supervised)*. At the end of February 2022, we update the review process. We used the same libraries and keywords in this second stage.

2.3 Articles selection procedure

We selected the returned articles by the search expression and read their titles, abstracts, and keywords. We read the method and experiments when we had doubts about the suitability of the article for the proposed survey. We reject the articles that meet at least one exclusion criterion. The exclusion criteria considered were:

- Publication date of the article was before the initial date of the search;
- Language of the article other than English;
- Systematic Review, Survey, and Chapter publication;
- Article without experiments;
- No access;
- Not suitable for the proposed objective;

2.4 Information extraction strategy

We did a full reading of the selected articles considering the following items that guided the information extraction:

- Title, publication year, country, library;
- Application domain;
- Objective;
- Dataset language;
- Text representation;
- Semi-supervised approach;
- Machine learning algorithms and/or deep learning method;
- Binary, and/or multi-class, and/or multi-label classification;
- Evaluation metrics;
- Classification results.

3 Results and discussion

This section presents the results obtained through the review. Figure 1 depicts the survey process. We achieved 1794 articles with the search strategy. From this group, 1637 articles were rejected according to the exclusion criteria. Thus, 157 articles were selected to perform a full reading and extract their information. Then, we performed a quantitative analysis to comprise the semi-supervised text classification information.

Following, Sect. 3.1 shows the number of publications per year and per country. Section 3.2 presents the main researched idioms, domains, and tasks. Section 3.3 shows the

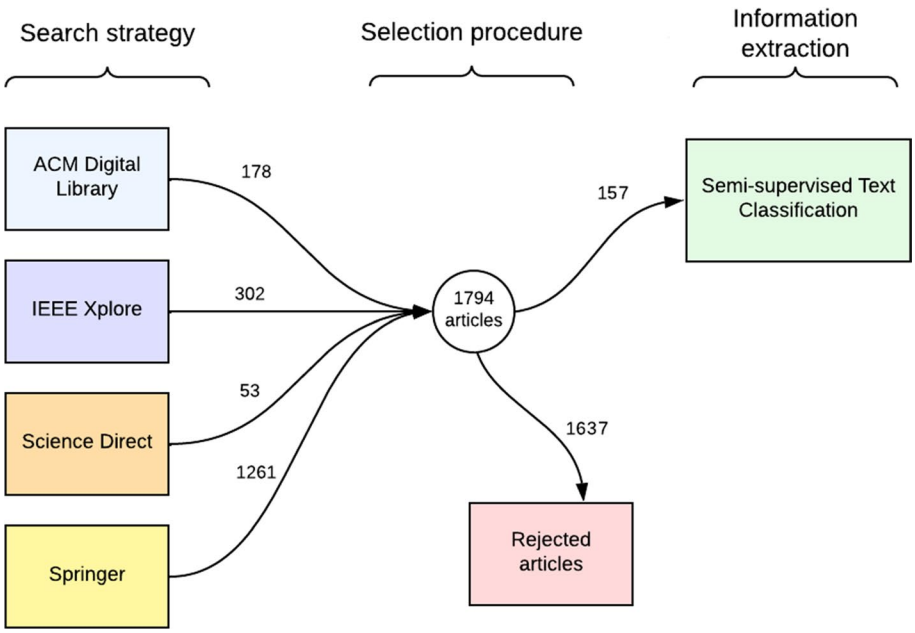

Fig. 1 Survey process to SSL for text classification. We access four digital libraries: ACM, IEEE Explore, Science Direct, and Springer. Applying the search strategy we retrieve 1794 articles. After applying the exclusion criteria 157 articles were selected to be included in this review

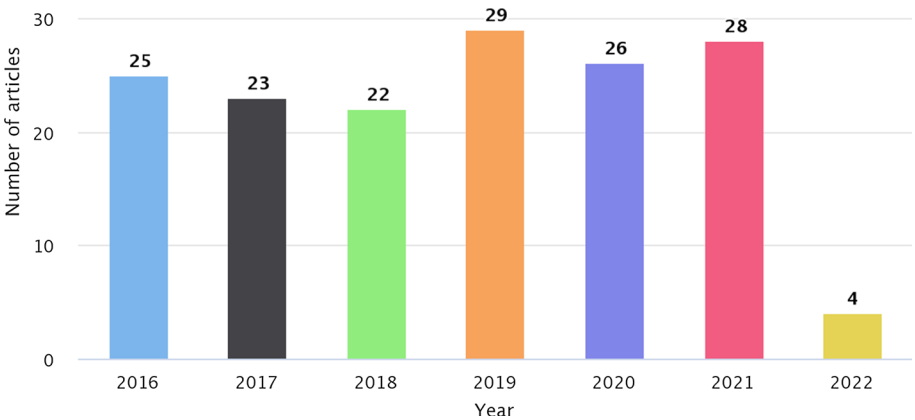

Fig. 2 Number of articles published by year

common datasets used. Section 3.4 presents the text representations and Sect. 3.5 the SSL approaches investigated.

3.1 Publications per year and per countries

Figure 2 depicts the scientific production per year. Since 2019 there has been an increase in the application of artificial neural networks (ANN) in the semi-supervised process. The

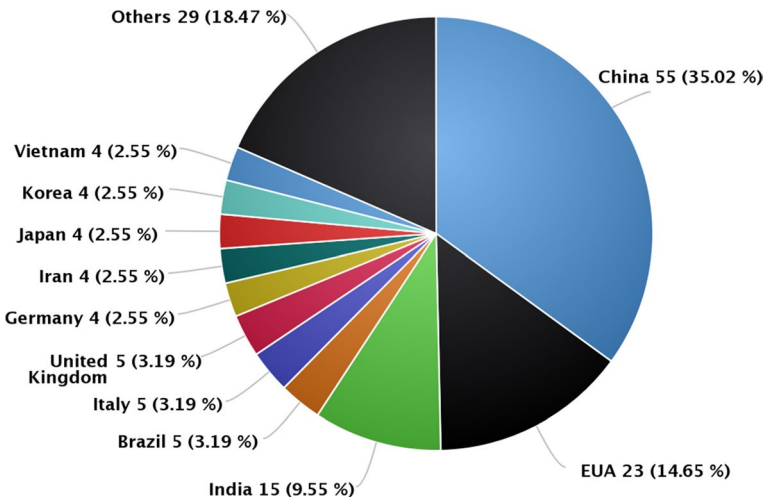

Fig. 3 Publications per country

years 2016, 2017, and 2018 had 17 publications that explored feature engineering or assist semi-supervised approaches using ANN. In 2019, 19 publications used ANN, 2020, 2021, and the first two months of 2022 had a total of 40 articles exploring ANN too.

We identify 33 countries that published semi-supervised text classification articles. Figure 3 shows the number of articles published per country, we included countries with at least four articles for visual and aesthetic reasons. China, the United States (USA), and India are the countries that most produced articles. China published 55 articles which represent 35.02% of the total of articles produced by all countries cataloged in this survey. After, we have the USA with 23 articles, and India with 15 articles, which represent 14.65%, and 9.55% of the total published articles, respectively.

Brazil, Italy, and United Kingdom published five articles each of them. Germany, Iran, Japan, Korea, and Vietnam with four published articles per country. Turkey published three articles, and the remaining 21 countries published one or two articles each of them. It is known that China has overtaken the United States and it is the world's largest producer of scientific articles. However, the USA is still considered a scientific powerhouse with high-level publications (Tollefson 2018).

3.2 Explored idioms, domains and tasks

Most of the NLP research was applied to English idioms, we identified 127 articles which correspond to 77.43% of the analyzed articles as shown in Fig. 4. Despite a small number of published articles, we identify 15 idioms other than English that investigated the text classification to their natural language. The Chinese had 18 (10.98%) published articles, Vietnamese 4 (2.44%), Arabic, Italian, and Brazilian Portuguese had 2 articles in each language. Each of the 11 remaining languages had one article published.

We distinguish 16 domains along with SSL applications. There are articles associated with more than one domain, then, the quantity of the articles distributed in the domains was 220, as shown in Fig. 5. We only present seven domains for viewing reasons. **News** was the most used domain in text classification with 56 (25.45%) articles. The majority of

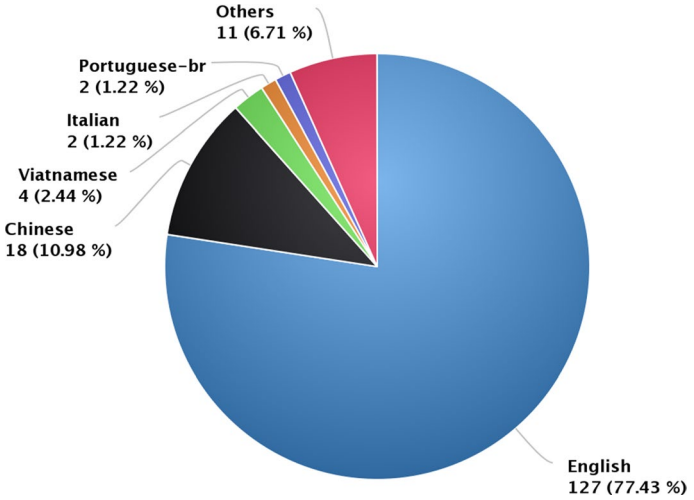

Fig. 4 Idioms explored in the papers

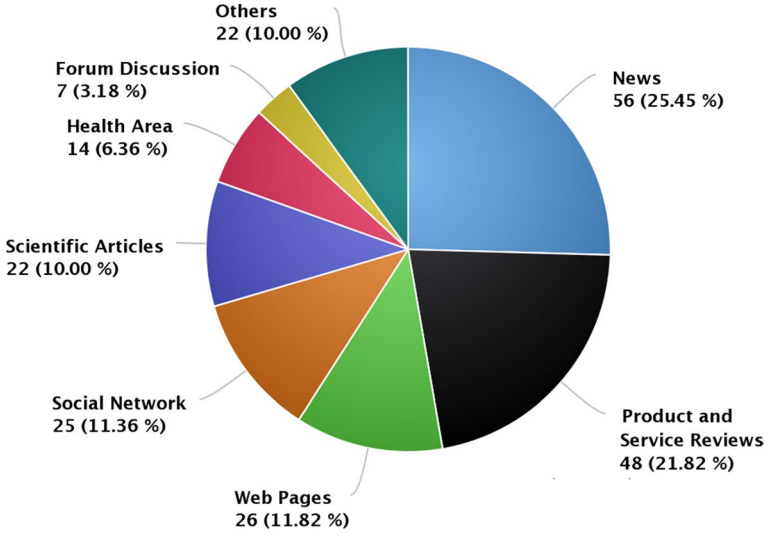

Fig. 5 Number of published articles per domain

News datasets are accessible benchmarks with known and verified outcomes, such as 20 Newsgroups, and Reuters 21578. In e-commerce, before customers make a purchase or hire a service, it is common for them to seek information from consumers about certain brands or services. Sentiment analysis supports e-commerce companies to understand the consumers feeling about their items for decision-making. We observed an increasing number of articles published in the **Product and Service Reviews** domain during the years analyzed, we count a total of 48 (21.82%) articles, where Amazon, Yelp, TripAdvisor, Movie Review, and IMDB were the prevalent datasets used.

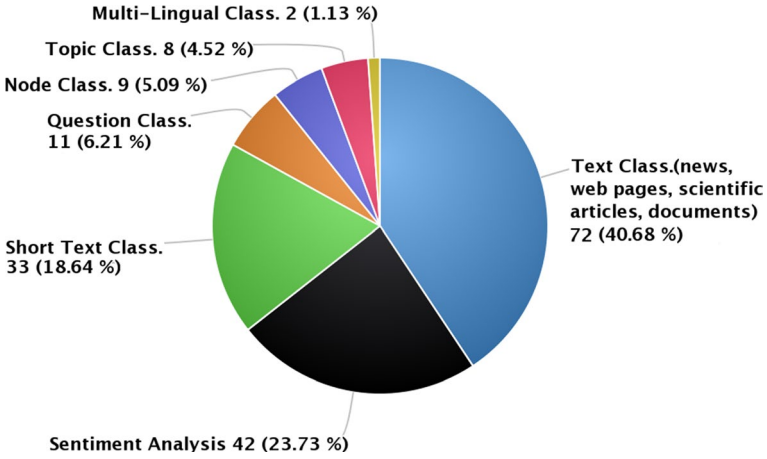

Fig. 6 Text classification tasks

Currently, there are approximately 400 million Twitter users in the world with 206 million active users per day (Dean 2022). Therefore, social networks generate abundant material that can be explored for the understanding of social behavior and its implications, e.g. sentiment analysis, emergency event detection, political purpose, fake news detection, and epidemiological studies. **Social Network** domain had 25 (11.36%) articles dealing with sentiment analysis or short text classification. **Forum Discussion** had 7 (3.18%) articles, the domain encompasses online discussions through a web platform where the users share their knowledge and argue about a determined topic, and the generated textual data can be applied to text classification tasks, i.e question classification.

In the **Web Pages** domain the articles mainly used WebKB or DBpedia datasets, and the total of articles was 26 (11.82%). **Scientific Articles** domain had 22 (10.00%) articles mainly related to the node classification. The **Health Area** domain with 14 (6.36%) articles and the Ohsumed dataset about medical abstracts was the most used. Different domains of the aforementioned had fewer articles: email, patent documents, internet advertisement, quotation, law, and education. Thus, these domains represented by **Others** had a total of 22 (10.00%) articles.

We organize the text classification into seven tasks according to Fig. 6. Some articles were applied in more than one task, then we had 177 articles distributed in the tasks. Generic **Text Classification** task was related to news, web pages, scientific articles, and documents that were explored in 72 (40.68%) articles. Then, **Sentiment Analysis** with 42 articles represents 23.73% of the total. The **Short Text Classification** task had 33 (18.64%) articles, in this case, we considered sentences, and microblogging when it is not used for the **Sentiment Analysis** task, e.g. sarcasm detection, intention detection, misinformation detection, rumor detection, irony detection, fake comments, and so on. The **Question, Node, Topic, and Multi-Lingual Classification** had 11 (6.21%), 9 (5.09%), 8 (4.52%), and 2 (1.13%) articles, respectively.

Table 1 Datasets used by the domain (Chinese -zh, and Vietnamese -vi)

Domains	Language	Datasets	Articles
News	en	20 Newsgroups	24
	en	Reuters 21578	14
	en	AG News	7
	en	Reuters RCV1	7
	en	Reuters RCV2	5
	zh	Sogou News	3
Social Network	en	Twitter	12
	zh	Weibo	4
Product and Service Review	en	Amazon	15
	en	IMDB	13
	en	Movie Review	8
	en	Yelp	7
	vi	Vietnamese Rev.	4
	en	TripAdvisor	3
Scientific and Medical	en	Cora	10
	en	CiteSeer	8
	en	PubMed	7
	en	Ohsumed	5
	en	Delve	2
	en	DBLP	2
Web Page	en	WebKB	7
Questions	en	TREC	6

3.3 Datasets

Table 1 represents the benchmark datasets most used in the experiments, totalizing 22 datasets. However, we identify other 114 datasets, but they are specific which makes them unfeasible for a possible comparison among the semi-supervised methods. The news domain had a total of 60 (34.68%) articles, where 20 Newsgroups and Reuters with 50 articles prevailed over AG News and Sogou News datasets, the last one is a Chinese benchmark. Concerning a short text, Social Network, and Product and Service Review domains had 66 (38.15%) articles that were used for sentiment analysis, social bot detection, deceptive review detection, and spam classification. In the Social Network domain, Twitter was quite explored in the English language with 12 articles, and Weibo in the Chinese language with 4 articles. In the Product and Service Review domain, Movie Review and IMDB dataset had 21 (12.14%) articles, and Amazon product categories (Books, DVD, Electronics, Kitchen, Music, Video) had 15 (8.67%) articles. Yelp, TripAdvisor, and Vietnamese datasets had a total of 14 articles, they are user reviews from restaurants, hotels, and places. The scientific and Medical domain includes 34 (19.65%) articles related to scientific publications and most of them were experienced with the Graph-based approach because the benchmark datasets structures are appropriate for node classification: CiteSeer, PubMed, and DBLP. On the Web Page domain, 7 (4.05%) articles used the WebKB dataset which is formed by web pages from

Table 2 The most used text representation

Method	Number of articles
TF–IDF/BoW/TF/IDF/N-gram	60
Deep neural word embedding	35
Word2Vec/Sent2Vec/Doc2Vec	22
BERT/DistilBERT/ALBERT	13
LDA/LSA	8
fastText	5
GloVe	4
Information Gain/Mutual Information	3
ELMo	2

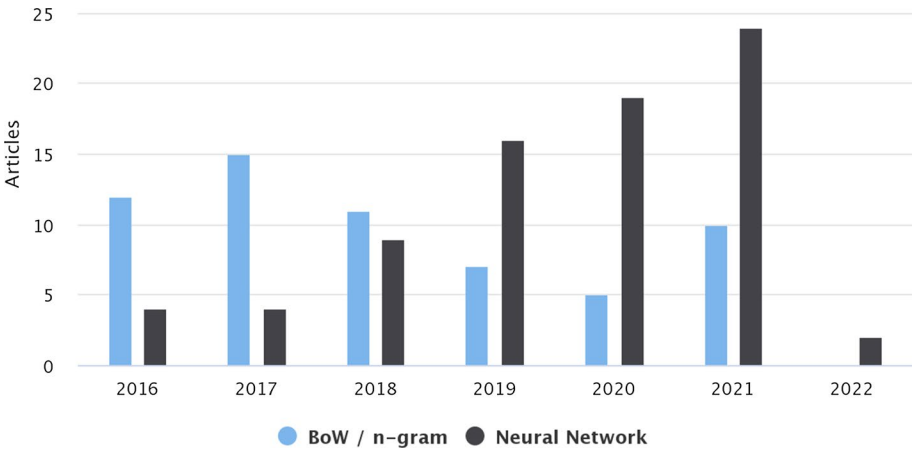

Fig. 7 Feature engineering methods based on neural network and term occurrence

computer science departments of various universities. Lastly, the TREC dataset with 6 (3.47%) articles for question classification.

3.4 Text representations

Table 2 displays the different types of text representation or feature engineering methods and their quantities applied in the text classification process in descending order. Despite bag of word (BoW), and term frequency–inverse document frequency (TF–IDF) being simplified methods, they are still quite used. Word2Vec, fastText, and GloVe are language models that handle lexical semantics, the first two are based on ANN, and the last one is based on word co-occurrence. Word2Vec and its extensions, e.g. Sent2Vec, and Doc2Vec had 22 articles related to them. FastText had five articles, and GloVe had four articles. BERT, DistilBERT, ALBERT, and ELMo are context-sensitive word embedding methods, we identify 13 articles referring to the first 3 methods and 2 articles to the last one. We identified 35 articles that implemented deep learning methods to generate or improve word

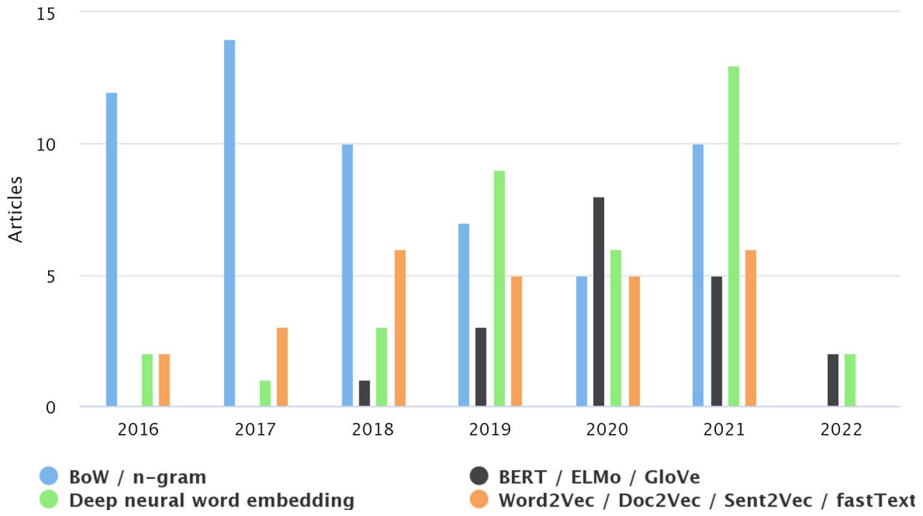

Fig. 8 Feature engineering methods per year

embeddings. Latent semantic analysis (LSA) and latent Dirichlet allocation (LDA) had eight articles. Information gain and mutual information had three articles.

A comparison with the feature engineering methods based on ANN and based on term occurrence/frequency is shown in Fig. 7.

ANN had an increasing application over the years with four articles in 2017 and nine in 2018, respectively. However, 2019 had a sharp increase with 16 articles, then 2020 with 19 articles, and 24 articles in 2021. The simplest methods of text representations had decreased in their use since 2017. Although, in 2021 the number of articles using traditional methods was doubled compared with 2020. In many cases, traditional text representation methods were used to do a comparison with contextualized vector representations or/and as input to an ANN.

Figure 8 exhibits in more detail the frequency of published articles over the years using ANN methods, and the methods based on term occurrence and term frequency. Word2Vec and its extensions had an increase between 2016 and 2018, but they remained practically stable in the following years. Context-sensitive pre-trained models applied in the experiments appeared in 2019 with BERT, and in 2020, and 2021 appeared experiments using ELMo too. Experiments with deep learning algorithms using their embedding layer had two articles in 2016, one article in 2017, three in 2018, and an expressive number of published articles in the following years. GloVe is based on co-occurrence matrices from Corpus and it is not context-sensitive such as Word2Vec and fastText. GloVe had one article in 2018, and 2019 and two articles in 2022.

3.5 Semi-supervised approaches

We followed the taxonomy proposed by van Engelen and Hoos (2019) to categorize semi-supervised approaches, as shown in Fig. 9. Meantime, we had the boldness to insert new approaches in the spectrum of semi-supervised algorithms to ensure the articles' categorization when the method did not match the taxonomy. Thus, considering the main method feature, we group the remaining articles in transfer learning and transductive support

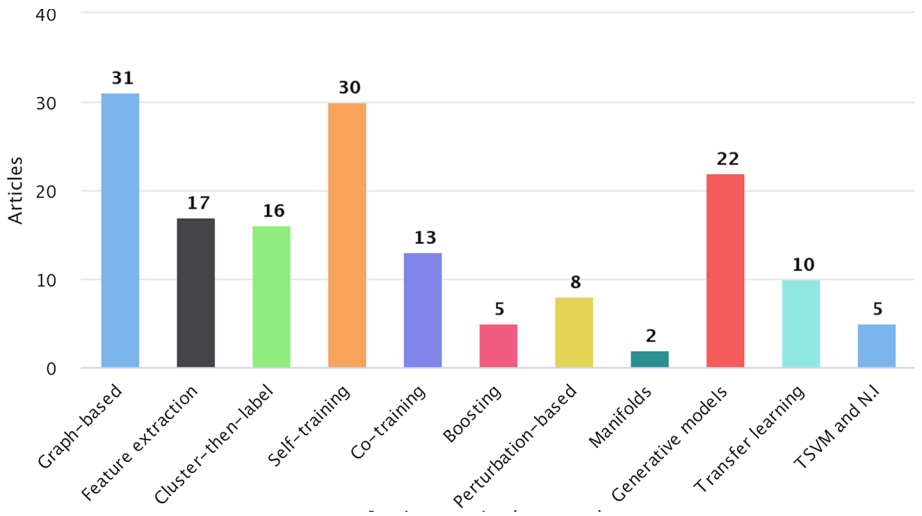

Fig. 9 Semi-supervised approaches per published articles

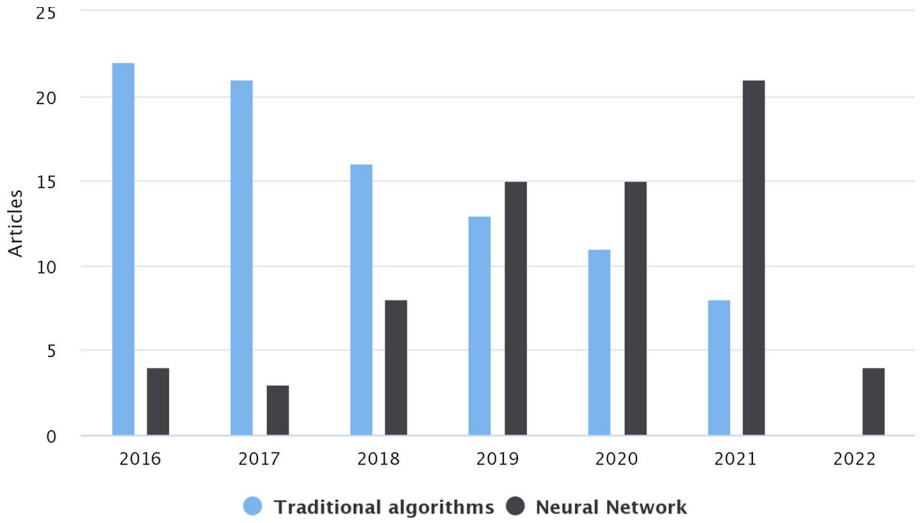

Fig. 10 Traditional algorithms versus neural network algorithms applied to the semi-supervised approaches

vector machines (TSVM) and not identification (N.I.) approaches. For the transfer learning approach, we separate articles that used jointly a limited number of labeled and a large amount of unlabeled target data in the training. Two articles were related to TSVM, three articles do not have a consensual opinion about the approach used.

With 31 articles the Graph-based was the most used technique. Until 2018, 15 of 17 articles with the graph method were not related to the ANN. Nevertheless, since 2019, 11 of 16 articles combined ANN and graphs. After 2019, 30 articles employed the Self-training approach, of these 21 articles applied traditional methods in feature engineering and text classification

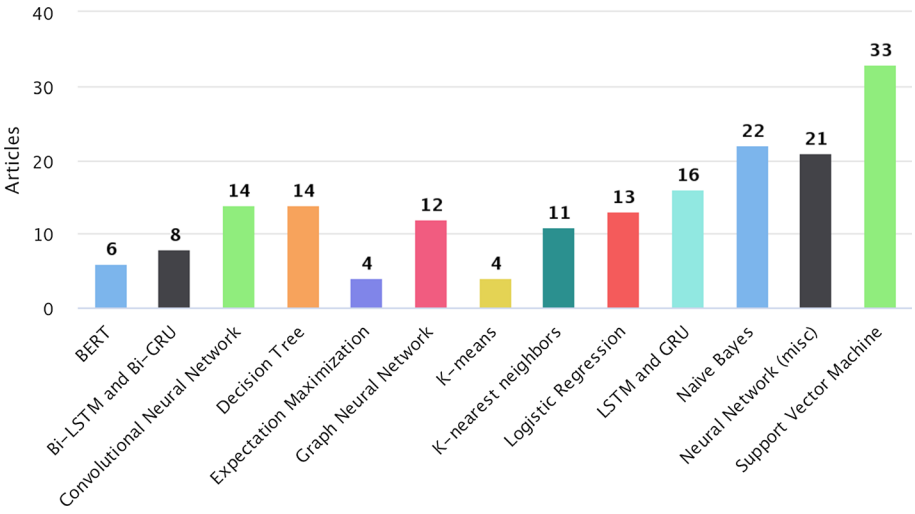

Fig. 11 Traditional and neural network algorithms

algorithms, and nine articles employed ANN. The third most used approach was Generative models with a total of 22 articles, in which 14 applied ANN. Then, Feature extraction, Cluster-then-label, Co-training, Transfer learning, Perturbation-based, Boosting, TSVM and N.I., and Manifolds with 17, 16, 13, 10, 8, 5, 5, and 2 articles, respectively. Without considering the first three most used approaches, there is a total of 76 articles in the remaining approaches which 35 articles applied ANN.

As can be seen in the previous paragraph, there has been an inclination to semi-supervised approaches using ANN over the years. Figure 10 clearly shows the behavior of traditional and ANN algorithms since 2016. There has been an increase in ANN and a decrease in traditional algorithms in the semi-supervised area. Although, the use of traditional algorithms has been shrinking, until 2018 it has a superiority compared to ANN. The year 2019 seems to be the inversion point, thenceforth ANN predominated in the semi-supervised approaches.

A comparison of the ANN and traditional algorithms applied in the semi-supervised approaches is shown in Fig. 11. Concerning articles with traditional algorithms, SVM frequency was 33 (18.54%), Naive Bayes with 22 (12.36%), and decision tree with 14 (7.87%). The decision tree includes CART, J48, random forest, and C4.5 algorithms. Then, Logistic Regression, k-nearest neighbors (kNN), K-Means, and EM algorithms with 13, 11, 4, and 4 articles, respectively.

ANN algorithms were grouped by their methods: long short-term memory (LSTM) and gated recurrent unit (GRU), Graph Neural Network (GNN), Convolutional Neural Network (CNN), bi-LSTM and bi-GRU, BERT, and neural network (misc), i.e. miscellaneous, but outnumbered algorithms. Neural network (misc) with 21 (11.80%) articles include different types of algorithms, e.g. multi-layer perceptron (MLP), autoencoder, ladder network, deep belief network (DBN), and capsule network. LSTM and GRU were used in 16 articles, while bi-LSTM and bi-GRU were used in 8 articles, and recurrent neural network (RNN)

algorithms comprise 24 (13.48%) articles. CNN, GNN, and BERT were applied in 14, 12, and 6, respectively.

4 Semi-supervised learning for text classification

In this section, we present the main works using SSL and text mining. We divide the topics following the taxonomy proposed by Van Engelen and Hoos (2020). Section 4.1 presents the graph-based approaches. Section 4.2 presents the unsupervised pre-processing approaches, especially the feature extraction and cluster-than-label methods. Section 4.3 presents the wrapper methods, especially self-training, co-training, and boosting. Section 4.4 presents the intrinsically SSL approaches, especially the perturbation-based, manifolds, and generative models. We also include the transfer learning methods in Sect. 4.5 and other approaches in Sect. 4.6.

4.1 Graph-based

Graph-based SSL methods propagate the labels to unlabeled nodes in a constructed graph $G = (V, E)$, where $V = \{V_l \cup V_u\}$ is a set formed by the labeled nodes V_l, and the set of unlabeled nodes V_u. V is a set of nodes, such that $V = \{v_1, v_2, \ldots, v_n\}$ represents the data points. E is related with a $n \times n$ matrix W containing for each pair of nodes v_i and v_j a non-negative edge weight $w_{i,j}$. The edge weight represents the similarity between the nodes.

Graph-based methods have been used in various contexts, e.g. news, web pages, health and scientific articles. We group the articles by context or text classification tasks to describe their methods. Regarding news classification, a method based on Positive and Unlabeled Learning (PUL) with Label Propagation (LP) to minimize the news labeling effort was proposed by Souza et al. (2021). Negative document extraction with graph paths based on Dijkstra's algorithm was proposed by Carnevali et al. (2021). They used sparse graphs for graph construction and Gaussian Field Harmonic Functions (GFHF), and Local and Global Consistency (LLGC) algorithms for classification. Authors in Yadav et al. (2019) compared distance/similarity metrics (Euclidean L2 norm; cosine similarity; improved sqrt-cosine similarity) to measure its effect on the quality of graph construction (Average Node Degree, and Standard Deviation of the Node Degree). The extraction of relevant content from the news web pages was carried out by Bose and Mukherjee (2019). The web page was represented as a graph, where text elements are nodes and the edge weights represent the similarity between nodes. A few nodes were labeled in the graph using heuristics and the remaining nodes were labeled by a weighted measure of similarity to the labeled nodes.

A graph-based algorithm to solve the label insufficiency by means of LP in the news dataset was studied by Gong et al. (2017). They explored two measures, i.e Graph Trend Filtering ad Smooth Eigenbase Pursuit to handle label inaccuracy by filtering out initial noisy labels. Widmann and Verberne (2017) constructed a graph employing documents nodes and features nodes where the order of the word was preserved. The connection was formed in two ways, i.e. among document nodes and features nodes, and features nodes based on words. A matrix representation of the graph was constructed with extracted features to LP based on context similarity using the Jaccard index. In a

multi-head-Pooling-based on Graph Convolutional Network (GCN) applied for news text classification, Zhao et al. (2022) focused on the structural information of the text graph for pre-training word embedding as the initial node feature. Important nodes were evaluated and selected from multiple perspectives through multi-head pooling.

In the news context, some works explored k-partite graph for text classification, where the vertices are partitioned into k different sets. A tripartite graph was developed by Ganiz (2016). Semantics in higher-order co-occurrence paths between words were exploited, which linked terms in unlabeled documents to terms in labeled documents. Furthermore, the method was able to estimate class conditional probabilities for the terms in unlabeled documents. Rossi et al. (2017) represented text collections by the bipartite heterogeneous network, where objects were documents and terms, and term and document were connected if there was a term occurrence in the document. The label of connected terms was propagated to a new document using a weighted linear function.

In a Chinese text classification for news, Zhu et al. (2018) developed a method based on Wikipedia sample extension (WSE). A network graph was constructed with concepts and their links extracted from Wikipedia. The generated extension was carried out by correlation of the labeled sample data and the concepts in Wikipedia by means of TF–IDF and then calculated the significant value of each concept for each category. Besides, to further expand the sample, was proposed WSEs with links (WSE-L), i.e. an enhanced sample extension method. After, it was placed a limiting condition to WSE-L to control the number of the training sample. Zhang et al. (2019a) investigated a news text classification based on a domain ontology graph of semi-supervised conceptual clustering. To deal with the problem of WSD, a framework of ontology learning of Chinese classification in accordance with the structural model of the domain ontology graph was developed.

Semi-supervised fake news detection method based on GNN was investigated by Benamira et al. (2019). GloVe of news articles was generated, and contextual similarities among texts were produced by kNN along with Euclidean distances in the embedding space. GCN and Attention GNN were used for the classification task. For the misinformation detection task, Abdali et al. (2021) studied three aspects of a news article which were combined and modeled as a tensor/matrix, with one model for each aspect. A hierarchical approach for finding latent patterns derived from those aspects was proposed. The nearest-neighbors graph was constructed with the articles in the embedding space for the semi-supervised label inference of unknown news articles.

Dealing with a short text classification task, Ji et al. (2021) proposed a streaming social traffic event detection via multiple edge computing based on heterogeneous information network (HIN) and clustering method. GNN along with HIN to obtain the optimal meta-path weights for traffic event detection was applied to measure the relationships between social texts. Binary sample GCN and binary sample graph-attention network (GAT) were constructed to address the problem of a large number of traffic event categories and a small number of samples in each category. Zhao et al. (2022), beyond news classification as described previously, applied the method for short text classification too. The smoothness assumption to the question of transductive multi-label learning was employed in Sun et al. (2018) with the purpose to exploit the correlation in the feature space and label space. A non-negative matrix factorization (NMF) based modeling and training algorithm which learned from adjacencies of the instances and labels from the training set was proposed. Employing a non-negative least square optimization algorithm, the labels were exploited and propagated.

In the short text classification task context, Kernel-based GNN for graph classification in social networks and movie reviews was investigated by Ju et al. (2022). Graph kernels were

combined with GNNs to effectively learn graph representation and used graph similarity for prediction. WordNet for WSD was used in Billal et al. (2017) and created a weakly connected graph through the words of the corpus with their synsets to extract connected components, where a component are nodes (words) and the edges are semantic relations among components. Furthermore, in a multi-label classification, semi-supervised graph methods were proposed for the extraction of subjects from the social network. Classification Maximization Deep Multi-label and Classification Maximization Deep Back Propagation Neural Networks were applied in the experiments.

For the short text classification task in Yang et al. (2021a), heterogeneous information embedding was carried out by heterogeneous GAT. Dual-level attention mechanism was applied to learn the weights and to capture the importance of different types of neighboring. Xu and Li (2017) developed a sentiment classification method based on a LP algorithm. The method combined text content and user information to construct similarity based on the reviewer's score preference and text features. High similarity between scoring preference and text features enabled to propagate of scores to unlabeled reviews.

Dealing with short text classification in a language other than English, Wang et al. (2017) performed a comparative study of algorithm performance with Chinese online reviews from multi-domains to resolve the problems of robustness and field dependence. Charalampakis et al. (2016) detected irony in a corpus of Greek political tweets researching training-collective classification. The goal was to find a relation between the ironic tweets that refer to the political parties and leaders in Greece in the pre-election period of May 2012, and their actual election results. Guo et al. (2016) focused on analyzing the credibility of influenza posts published on Sina Weibo by means of user, content, and post. An undirected Graph Markov Network with random variables was used to model dependencies among nodes and to capture interactions among features.

In the scientific context, considering the importance of the external information of nodes to improve the performance of representation learning, Liu et al. (2018a) applied the Hierarchical Attention Network Embedding method which performed integration between text and label features of nodes to learn the hierarchical relational network embeddings for scientific articles. Two layers of bi-GRU were applied to the hierarchical learning: one layer extracted latent features of words with word-level attention to obtain the lexical features, and the other one extracted latent features of sentences with sentence-level attention to obtain textual features. Zhu et al. (2021) researched random walk and GNN using global and local information to handle scientific articles. Global information was preserved by global features. A set of parallel kernel GNNs was used to learn different aspects of pretrained global features and the raw attributes of the graph. Yang et al. (2021b) explored multilayer GCNs to handle complexity and redundant calculations, and the overfitting problem of GCNs. A simplified multilayer GCN with dropout which extends shallow GCNs was applied in scientific texts.

In scientific texts, Xu et al. (2020) investigated label consistency with GNN that generated label distribution for each node in addition to the similarity to aggregation weight between two nodes. The method benefited from the proportion of neighboring nodes with the same label, and of the target nodes and unconnected nodes that shared the same labels. Akujuobi et al. (2020) applied recurrent-attention strategy to handle the problem of a large number of neighboring nodes to be analyzed and used inductive properties in semi-supervised node classification using scientific articles. The walk on the graph was learned based on recurrent attention which reduced the noise information, interpreted the decision-making process, and inferred class label dependency. GAT to label propagation was applied by Huang et al. (2021), and the graph was constructed considering citation datasets properties.

The embedding vector of each node was generated based on their neighborhood. An attention mechanism was applied to learn the representation of neighbor nodes of target nodes, then nodes with high similarity to target nodes had higher weights, and low similarity nodes had lower weights.

For scientific text classification, a dynamic anchor graph to learn local and global features jointly was elaborated by Wang et al. (2021). A two-branch architecture was built, one branch was single-sample consistency that learned local features by consistency regularization term, and the other one used outputs from the previous branch to construct dynamically an anchor graph. Graph embedding branch learned global features in the graph by context prediction log loss. Timsina et al. (2016) investigated various SSL including label-spreading along with Radial Basis Function kernel to select articles for medical systematic reviews. In Kontonatsios et al. (2017), an active learning method was proposed to contribute to citation screening in clinical and public health reviews. The approach was based on cluster assumption and used label propagation to neighboring unlabeled citations supported by cosine similarity measure applied in the feature space.

4.2 Unsupervised pre-processing

4.2.1 Feature extraction

Unsupervised preprocessing is a category of inductive methods that use unlabeled and labeled data in dissociated stages, where the unsupervised stage can be done by feature extraction. In NLP, feature extraction converts the raw text data into numeric features which are able to improve the performance of the classifier. Feature extraction is an SSL method carried out on unlabeled data and seeks to extract relevant information from the raw data, and it uses a supervised fine-tuning stage (van Engelen and Hoos 2019).

In the news context, using CNN for multi-label classification, Li et al. (2018) presented the following process: words were extracted from legal documents, and Word2Vec generated the word embeddings; two view embedding learning generated training data; predicted target regions with feature regions by training; two view embeddings were integrated into CNN for text classification. Jiang et al. (2018) combined DBN and Softmax Regression forming a hybrid algorithm, where the features were learned by DBN, and softmax regression was trained along with a few samples labeled. In the fine-tuning step, the system parameters were optimized with limited memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) algorithm which used estimation to the inverse Hessian matrix and cost function with second-order Taylor expansion.

With an approach to the sentiment analysis task, Pan et al. (2020) used Ladder Network that integrated a small amount of labeled data with a large number of unlabeled reviews and augmented data effectively. The method has two models, the first one leveraging contextual features from unlabeled data using either Word2Vec, BERT, DistilBERT, or ALBERT. The second model was Ladder Network along with an encoder and decoder model. For sentiment classification, Sun et al. (2019a) explored fine-tuning methods of BERT. The within-task and in-domain further pre-training boosted text classification performance and improved the task with small-size data. For affect classification, Chawla et al. (2019) introduced a deep neural network in an environment with limited labeled data, the method was a gated sequence-to-sequence, convolutional–deconvolutional auto-encoding. The classification of tweets was addressed in Baecchi et al. (2015), according to their polarity, considering both textual and visual information. A novel schema was proposed incorporating a

CBOW with negative sampling and Denoising Auto-encoders to exploit web-scale sources corpus and robust visual features obtained from unsupervised learning.

For sentiment analysis tasks in languages other than English, Jahanbakhsh et al. (2020) researched a model based on content and context features to verify Persian rumors. The content-based features were a set of writing style features, and the context-based features were speech acts of rumor documents and contextual word embeddings that were extracted by two parallel BERT models. In Guellil et al. (2021), an approach for sentiment analysis of Arabic messages extracted from social media was proposed. Both Arabic and Arabizi (it used Arabizi transliteration and Arabizi translation to Arabic) were considered by the method and Word-2Vec associated with the classical algorithms was applied. Besides, Word2Vec and fastText plus deep learning Algorithms (CNN, LSTM, bi-LSTM) were applied too. Di Capua and Petrosino (2017) experimented with the ANN model based on DBN which learned feature representations from labeled and unlabeled data. The method was built to deal with data uncertainty for sentiment analysis and adopted the Italian language. Yadav and Bhojane (2019) developed an SSL approach to sentiment analysis in Hindi language documents. The authors worked with three approaches: ANN with pre-classified words; classification using Hindi SentiWordNet; classification with ANN and pre-classified sentences.

This paragraph summarizes text classification in Japanese and Chinese languages. For the Japanese language, automatic section identification of requests for quotation documents was developed. Novel features were introduced derived from unlabeled data to enhance the performance, e.g. lexicon features, word cluster features with Word2Vec, and cluster features with constraints (Hidetaka and Wang 2019). For Chinese charge prediction, He et al. (2019) elaborated a Sequence Enhanced Capsule Model constituted by: an input layer where the words of fact description of a case were transformed to the primary capsule; multiple seq-caps layers, one layer produced advanced semantic representation from fact description and other one restored the sequence information of fact description; mechanism attention, a new residual unit improved the generalization and provided auxiliary information for charge prediction; the output layer, all the features vectors from the multiple seq-caps layers were flattened and concatenated with the global context vector, then the fully connected network and softmax function were used to generate the probability.

In a web page domain, Geraci and Papini (2018) built automatically a set of examples to use as the training set. The method exploited the strong correlation between URLs text representation and text from the web page, therefore a set of web pages per class was constructed. Vectors of features were built per class/URL pair and were used to label URLs by ranking the classes. In McNulty et al. (2021), an approach that classifies HTML documents in research and non-research based on structural, content, and formality features was explored. In Lieder et al. (2019), millions of public business web pages were mined and they used multi-lingual BERT to obtain a contextualized representation of texts and CNN multi-label text classification. Due to the fact that missing labels affect the classification performance for multi-label learning, Cheng et al. (2021) approached missing multi-label learning with non-equilibrium based on a two-level autoencoder to web page classification. Two-level auto-encoder was constructed considering the noise interference in the feature space and the correlation between features and labels.

For sentence classification, the fastText model was analyzed by Agibetov et al. (2018) in biomedical sentences. SSL models were pre-trained through unsupervised training on predicting word contexts or sentence reconstruction tasks and then used downstream supervised classification.

4.2.2 Cluster-then-label

It is an inductive method that uses unlabeled and labeled data in two stages, such as feature extraction. The unsupervised stage comprises the clustering of the data.

In the news context, Jedrzejowicz and Zakrzewska (2020) proposed a hybrid approach by the LDA algorithm and Word2Vec. The method clustered the documents into categories using topics in an unsupervised way. The results of collapsed Gibbs sampling for LDA were acquired and each topic was expanded by the word embeddings, similar to the most representative words from the topic, through the cosine distance metric. For a new document was calculated word-topic distribution for each word of the document, then the topic was assigned to the document which had the highest number of word-topic assignments. For news classification, Barman and Chowdhury (2018) used Kohonen's self-organizing map to extract the groups from texts and unlabeled samples of each group were labeled based on the voting of the class label with labeled members of the group. New classes were detected during the clustering process to news text categorization in Guru et al. (2016). Samples too far apart from all clusters in the clustering process formed one or more news clusters. The new cluster fully formed by unlabeled samples represented the new class, therefore the samples were labeled.

For online news article classification, Krishnamoorthy et al. (2018) used two incremental clustering methods. The method I calculated for each new document its cosine similarity with all of the original documents. Method II used the centroids of the original clusters rather than all the data points when calculating the cosine similarity values of the new document with the clusters. A selective seeding technique to obtain a coherent set of initial centroids based on maximum feature coverage was implemented. Vilhagra et al. (2020) elaborated a deep clustering approach to document clustering and feature learning through the K-Means algorithm, convolutional Siamese network (CSN), and pairwise constraints (cannot-link constraint, and must-link constraint). The CSN and pairwise constraints were used to learn a low-dimensional representation, the feature vectors were conducted by L_1 norm that brings them closer or farther away by semantic distance.

Still, in the news domain, Thomas and Resmipriya (2016) formed clusters with samples with the same labels, and they were identified by their centroids and labels. The distance between the unlabeled samples and the centroids of the labeled clusters was calculated, where the minimum distance defined the cluster target to the unlabeled sample to be added and labeled. The similarity metrics were Euclidean distance, cosine similarity measure, similarity measure for text processing (SMTP), and dice coefficient. For news classification, an unbiased semi-supervised cluster (SSC) tree was proposed by Sun et al. (2020), in which the learning process used only very few labeled data, and a confidence error-based pruning algorithm. The K-Means algorithm was applied to generate the SSC tree, where each level of this hierarchical tree was built in a top-down manner, and the confidence error was used to prune the tree. With a global strategy based on the weak cluster assumption to explore the unlabeled data, the method proposed resolved the local maxima problem.

For the short text classification task, Ng and Carley (2021) examined coronavirus-related fact-checked stories. In K-Means clustering six topics were chosen, and each story was assigned to a cluster number based on its Euclidean distance to the cluster center in the projected space. BoW classifier was constructed to label the story type by means of cosine distance, and the BERT classifier to label the target story using the closest vector embedding found through the smallest cosine distance. Buza and Revina (2020) improved the

classification of time series and applied it to short text classification. Previously, labeled and unlabeled samples were clustered with constrained single-linkage hierarchical agglomerative clustering. Then, in the top-level clusters generated, the unlabeled samples in each cluster were labeled by their seeds. However, the complexity of distance computations was $\mathcal{O}(n^2)$. Considering the distance computations used in the old method, when the dataset was divided into parts (c) and computed *m-times*, the complexity became $\mathcal{O}\left(\frac{n^2}{c}\right)$. Therefore, the authors relied on this logic to reduce the computational cost.

A short text classification method based on weighted word vectors representation was proposed by Zhang et al. (2019b). Expected cross-entropy was used in the labeled data to extract strong category feature sets. To reduce the high-dimensional sparseness of features from short texts, word vectors were generated and used to represent eigenvectors increasing the semantic information of short texts. The method calculated the cosine similarity of the whole eigenvectors and the virtual class center, where the virtual class center represented the mean value of the eigenvectors. The real class center of the labeled samples was calculated based on normalized similarity. The similarity between the clustering center and the real class center of the labeled data was used to classify the unlabeled samples.

In the social media sentiment analysis task, Nguyen (2016) exploited the concept of emotional consistency with spectral-based LP and distant supervision labels or noisy labels. The LP was based on a similarity matrix that used a Gaussian kernel based on textual features. In the emotional clustering, consistency was built on three different predictors based on three lexicon resources using the lexicon-ratio method. The final sentiment classifier was built by the reference predictions and the labeled data of the target domain. Namrutha Sridhar et al. (2020) identified and associated social media text with multiple emotions with varying degrees. Word embeddings were trained for the entire Twitter dataset, then Twitter level similarity was calculated between unlabeled and labeled tweets by word mover's distance.

Two researchers performed short text classification in the Vietnamese language. First, Ha et al. (2018a) did a recursive adaptation multi-label classification algorithm with semi-supervised clustering. The method finds the first label (λ) as the greater number of occurrences in L_2 which is the set of possible labels that the labeled dataset might have. The clusters were created based on λ and generated three macro labels λ_1, λ_2, and λ_3 as simulated label set. A set of clusters (D_1, D_2, and D_3) related with the labels λ_1, λ_2, λ_3 was produced. Second, Ha et al. (2018b) proposed a lifelong topic modeling method, which focused on learning bias on the domain level based on the proposed domain closeness measure, and an application framework for multi-label classification based on semi-supervised clustering to Vietnamese texts.

In the scientific classification domain, Varghese et al. (2018) employed an unsupervised clustering algorithm with a minimal training dataset to cluster the labeling process to reduce the manual effort in the process of a systematic review of toxicological studies.

4.3 Wrapper methods

4.3.1 Self-training

Self-training approach is part of wrapper methods, whose logic of such methods is to generate pseudo-labels to unlabeled data, and add the additional labeled data generated along with the existing labeled data to train an inductive classifier.

In the news context, a modification in the self-training method was performed to reduce the sensitivity of the learning algorithm to the noise contained in the labeled data by means of automatically generated summaries by Villatoro-Tello et al. (2016). Another contribution to the research was a new strategy based on the distance to select confidently labeled instances in every iteration of self-training, which helped to preserve high homogeneity values among classes. In Pavlinek and Podgorelec (2017), the topic model to represent text was investigated with the aim to improve performance in the SSL method. The news text classification method based on self-training and LDA topic models was proposed to augment very small labeled data sets with unlabeled content. In Kumar et al. (2021) a novel framework of binary classifiers eliminated the threshold issue to improve the performance of pseudo labeling in the conventional SSL for text classification using a new dataset.

Dealing with news and sentiment context, a new hybrid method was built for classification which used class-based meaning values and weights of terms (Altınel and Ganiz 2016). The meanings of the words for the class were calculated and the meaning score defined the labels to unlabeled samples. After that, Class Weighting Kernel constructed the class-based matrix which represented the weights of the words for each class. Then, based on a class-based matrix a symmetric term-by-term semantic smoothing matrix was generated to calculate the similarity/kernel between documents. The kernel function was embedded into the implementation of the SVM algorithm used along with Platt's Sequential Minimal Optimization classifier. Still, for news and sentiment classification, Altnel et al. (2017) along with a meaning calculation computed the words' mean scores in the scope of classes. Instance labeling used meaning calculation in a semi-supervised way to construct a semantic smoothing kernel for SVM.

In the sentiment analysis task context, Khan et al. (2017) incorporated machine learning along with sentiment lexicon in order to alleviate existing problems of data unavailability, data sparsity and domain dependence. The sentiment knowledge base was constructed resulting in two sentiment lexicons named Senti-IG and Senti-Cosine by the application of mathematical models such as Information Gain and Cosine Similarity for the SentiWordNet lexicon to generate revised sentiment scores. A system was developed by Zaghdoudi and Glomann (2021) to automate user research activities on the web. The synonym replacement method was used for data augmentation, and LSTM was applied to sentiment analysis. For the sentiment and topic classification, Xiang and Yin (2021) combined deep neural network bi-GRU and temporal ensembling extended which unlabeled samples were labeled with pseudo labels. A sarcasm-unlabeled method was proposed by Li et al. (2020) for contextual sarcasm detection in social networks using the concatenation of content representation based on CNN and sarcastic preference embedding along with the main-balanced and main-unbalanced dataset.

SSL to sentiment classification as a model-based reinforcement learning problem was inspired by self-training in Li and Ye (2018). An adversarial network-based framework was proposed, but unlike most of the other generative adversarial network (GAN)-based SSL approaches, the framework did not need to reconstruct input data and hence could be applied for semi-supervised text classification. In Banerjee et al. (2018), sentiment classification was handled through positive and unlabeled data, when the positive class was a rare event in customer reviews. Stage I sought to label data for Non-Reportable and new kinds of Reportable cases and estimated the prior class probabilities by means of sentiment score, keyword score, and similarity score (using LSA or GloVe embeddings). Stage II used an entropy-regularized logistic classifier that penalized the entropy of the posterior measured on the unlabeled samples.

In Lee and Kim (2017), the sentiment labeling method was explored to generate confidently pseudo-labeled samples with threshold parameter which was added to the training corpus in order to enrich the initial sentiment classifier. In each iteration, the self-training with concatenated embedding vectors was conducted. Four experiments were carried out: sentiment classification to prove the effect of sentiment labeling; an experiment conducted to identify whether sentiment labeling with a lower confidence threshold could improve classification accuracy and to determine whether there was a correlation between the joint sentiment/topic model variance and classification accuracy; one experiment for further validation; with an increment of the size of initial-human-labeled data, an experiment was carried out to analyze the performance of the proposed method.

In the short text context, Shulman and Simo (2021) proposed a method based on deep learning for helping users in online social networks avoid regrettable posts and disclosure of sensitive information. A semi-supervised self-training approach was employed to incrementally label messages from online social networks and create a large-scale corpus. Word2Vec and fastText were used to generate domain-specific word embeddings. User information to alleviate the data sparsity in sentence classification in social scenarios was used by Ma et al. (2020). The up-based regularization term was applied to assist the prediction and in the self-training, the pseudo-labeled had noise reduction by a sample selector. A pre-trained ELMo was used to contextualize word embeddings and the softmax layer to output the probability distribution over classes. Deocadez et al. (2017) applied algorithms in order to automate the classification of functional and non-functional requirements contained in the App Store reviews.

For short text, the label prediction method was proposed by Stanojevic et al. (2019) which predicted probabilities to guide the choice of labels for each post from unlabeled data based on the small number of labeled samples. The method captured additional contexts from the unlabeled data with model learning, e.g. fastText, and deep learning models. With SSL framework for short text, Ghosh and Desarkar (2020) improved the performance of the classifier trained in a small labeled set incorporating highly confident samples from unlabeled data for labeled training data. One criterion for the class assignment and selection of samples was the restriction in the number of samples per class, and the other one was based on the class-specific threshold, which restricted the assignment of samples to class.

For short text classification, Karisani and Karisani (2021) proposed a neural SSL model based on a classic self-training algorithm that was threshold-free to cope with social network data. The method handled the semantic drift problem and revised the previously labeled documents. The approach was iterative and formed by two neural network classifiers that reverse each other. In each iteration, one classifier obtained a random set of unlabeled documents and labels them. This set was used to initialize the other classifier, to be further trained by the set of labeled documents. Three semi-supervised methods to classify tickets in a binary fashion from bug tracking system data were employed in Pohl et al. (2020), and sentiment polarities were used as a feature of the Self-training. Wulan and Supangkat (2017) proposed a semi-supervised Self-training to classify motivational messages which may motivate the learner to study.

In the context of languages other than English, Duong and Anh (2021) used Easy Data Augmentation, e.g. synonym replacement, random swap, random insert, and random delete to sentiment analysis in Vietnamese texts. Besides, syntax-Tree transformation and back translation data augmentation techniques. For sentiment analysis, Nguyen Nhat Dang and Duong (2019) has taken various experiments including many pre-processing techniques, and semantic lexicon complementation. Furthermore, synonym replacement and random

swap data augmentation techniques improved the accuracies of classifiers. Yin et al. (2018) applied the SSL method, SVM classifier (SLAS), and CART model for sentiment classification.

In Li et al. (2017a), a lot of unlabeled samples in the data set were labeled iteratively based on the similarity between the samples. A novel semi-supervised Chinese short text classification algorithm based on fusion similarity and class center was developed. Khan and Zubair (2020) proposed a model for the multi-lingual (English and Roman Urdu) classification of tweets into a multi-class model. The SSL method was based on a feature set from the labeled dataset, the unlabeled samples were labeled and the model was re-trained with them jointly and the smallest previous labeled set. Omar et al. (2021) focused on the short text classification on the social network and constructed a standard Arabic dataset using manual annotation and semi-supervised annotation techniques. One of several experiments was self-training used to label the remaining unlabeled posts with sentiment class.

In the health domain, a comparative analysis was performed on various SSL methods with the purpose to address the problem of the small training dataset to text classification algorithms in medical systematic review (Liu et al. 2018b). Self-training with label spreading to identify the most confident unlabeled instances was one of the semi-supervised methods used. Hasan et al. (2020) identified adverse drug reactions and side effects from a patient report on social media along with a semi-supervised method. The method was based on a Conditional Random Field with a small labeled dataset which iteratively augmented the training set with high-confidence labeled sentences coming from a large set of unlabeled data. Furthermore, incrementally the method augmented symptoms and side-effect dictionaries with the most confident medical terms. Thus, with the terms correctly classified, sentences that were rejected before could be added to the training data.

In the Web Page context, Lin et al. (2017) elaborated a competitive perspective identification based on user-level perspective consistency which selected high-quality classified texts from the unlabeled corpus and iteratively boosted the classifier. The method refined the perspective classifiers with the document-topic distributions mined from texts using NMF. SSL multi-view similarity for web page classification was designed by Wu et al. (2019). The method learned multiple view-individual transformations and one shareable transformation. Therefore, the particularity and commonality of different views were explored. Label information of labeled samples and the similarity information of unlabeled samples were used from both intra-view and inter-view aspects. The overall objective was given by the combination of the terms of semi-supervised multi-view similarity preserving, multi-view statistical uncorrelated design (to reduce information across views to learn view-specific features with view-individual transformation using covariance matrix), and classification loss. The $l_{2,1}$-*norm* base regularizer was employed for view-specific transformations that were sparse in rows, then discriminant features could be selected for each view.

4.3.2 Co-training

The approach is a semi-supervised method and a part of wrapper methods that use supervised algorithms to iteratively label unlabeled samples. The characterization of Co-training is given by the use of two or more distinct views of the labeled data to iteratively train the classifiers. At each iteration, the most confident prediction from each classifier is passed to the labeled data of the other classifiers.

In the news context, a collaborative text classification was combined with a supervised topic model to identify the semantic relation between the topic and category by Zhang et al. (2021a). The views were generated by different feature representations for training two classifiers, and the approach adopted a confidence calculation method based on posterior distribution distance and sampling strategy to select credible unlabeled samples. Xu et al. (2016) dealt with weakly labeled learning problems with multi-view training data, where pseudo-label vectors were used to pass information among different views. A projection operator was proposed, which converted the predictions to pseudo-label vectors considering different constraints in weakly labeled data from different learning scenarios. Multi-view semi-supervised co-training algorithm to news text classification was applied by Iglesias et al. (2016), where a BoW view and a new view from the BoW based on hidden Markov models (HMMs) were generated. A document group was constructed for each label and HMMs represented the groups. The classification of a new document was given by maximum probability value after the probability analysis of the document being generated by each of the HMMs.

For sentiment analysis, a new hybrid approach that combined context-dependent embeddings based on the ELMo language model along with co-training in an integrated perspective was investigated. The classification was carried out in an online social network of a German direct banking institution by Graef (2021). An adaptation was done by Alnashwan et al. (2019) in the co-training method to a multi-class classification to sentiment analysis in online medical forums about Lyme Disease, and Lupus.

In the scope of the question and short text classification task. Drug treatment question classification task using the co-training method in medical forums by bi-LSTM and bi-GRU was explored by Wang and Ren (2019). Random subspace method for co-training (RASCO) and relevant random subspace co-training (Rel-RASCO) to automate the classification in App Store reviews were applied by Deocadez et al. (2017). RASCO did random feature splits, while Rel-RASCO was a result of RASCO modification that changed random feature subspace ideas, and searched to select relevant feature sub-spaces. A novel design for CNN in SSL short text classification was presented by Shayegh et al. (2019). The dataset was partitioned into independent views via topic modeling to train independent classifiers. The kNN grouped views into unique categories based on their topic similarity to auxiliary classifiers to predict the label of documents. The method leveraged Words' synonyms to augment the dataset in addition to the original labeled training. A novel framework for learning from the text-rich network was proposed by Zhang et al. (2021b). With co-training algorithm and feature sharing, two modules were trained jointly, a text analysis module for text embedding by BERT, and a GNN module for categorical information propagation. The GNN model used neighborhood sampling and attention-based aggregation, the two modules had different inductive biases. SSL was applied in Jing (2018) for online fake comment detecting with dynamic and static features representations as views.

With the web page dataset, Gokhale and Fasli (2017) proposed a co-training SSL approach to the multi-class recognition problem to classify human rights abuses. A multi-labeled deep method that combined two-view for text classification by implementing two deep neural networks was proposed by Kihlman and Fasli (2021) to classify human rights violations. The method added noise data to the classifiers to learn to differentiate noise data and correct data, and so improve classification accuracy.

In the scientific context, the view insufficiency problem was addressed in Guo (2018), the method sought to identify harmful data and modify them, reducing their effects, i.e. decreasing their weights in the training set to scientific classification.

4.3.3 Boosting

In pseudo-labeled boosting methods, the classifier ensemble is formed by dependents base learners. The method trains models with supervised base learners using unlabeled samples, in each learning iteration the method generates pseudo-labeled which are incorporated with the labeled samples. Furthermore, in each learning iteration, the models are combined to build a single classification model (van Engelen and Hoos 2019).

In the news context, Tanha (2019) investigated a new multiclass loss function using new codewords to address the multiclass semi-supervised text classification problem. In the multiclass loss function, one term was the margin cost of the labeled data and the other one was a regularization term of the unlabeled data. In order to guide the base learning for assigning the pseudo-label to the unlabeled data, the loss function combined the pairwise similarity and the classifier predictions. A set of new different similarity functions was applied to improve the classification performance using different distance/metric learning algorithms, and boosting frameworks to derive an algorithm from the proposed loss function.

A new form of boosting framework for learning optimal similarity function to multiclass news text classification problem was proposed by Tanha (2018). The method combined the similarity information between labeled and unlabeled data with classifier predictions to assign pseudo-label for unlabeled examples. Based on cluster assumption and maximizing margin approach for multiclass case, a new risk function to multiclass semi-supervised classification problem was introduced. Weights were assigned to all data points which were used to find a new optimal classifier and decrease the risk function and used boosting framework to learn weak similarity functions. The final classification model was formed by a combination of weak classifiers and similarity functions. For news classification, Liu et al. (2016) elaborated an extension of the AdaBoost with Universum examples, where the training error was bounded by the product of the normalization factor.

In the sentiment analysis task, auto-labeled unlabeled tweets gathered by location from the USA along with emoticons to generate the training data were proposed by Hanafy et al. (2018). Features were extracted from labeled data by statistical and unsupervised approaches, e.g. TF–IDF and Word2Vec, respectively. Classical (SVM, MaxEnt) and deep learning methods (LSTM, CNN) were combined generating a unified model.

In languages other than English, Li et al. (2017b) employed an ensemble classifier based on Bagging and AdaBoost methods for Chinese question classification. A simple data editing technology based on kNN was applied for not to prejudice the classification model with predicted error labels from unlabeled samples. TF–IDF and lexical-semantic extension methods derived from Tongyici Cilin were used with Naive Bayes, J48graft, and J48 classifiers. The semantic extension method was compared with TF–IDF in supervised and semi-supervised methods.

4.4 Intrinsically semi-supervised

4.4.1 Perturbation-based

Intrinsically semi-supervised methods add unlabeled samples to the objective function, and they perform a direct objective function optimization. These methods modify the objective functions to include unlabeled data, thus they are considered enlargement of supervised

methods and do not depend on supervised base learners. Another feature of these methods is their dependence on one of the SSL assumptions. The maximum margin depends on the low-density assumption, and the decision boundary must remain in a low-density area, while the Perturbation-based method directly incorporates the smoothness assumption (van Engelen and Hoos 2019).

Local perturbations generate adversarial examples which are the results of imperceptible changes in samples. Prediction model robustness to local perturbations is a presupposed smoothness assumption. Thus, the predictions for imperceptible changes or noise in the sample and the unchanged sample should be similar. Unlabeled samples can be used because the similarity is not dependent on the true labels of the samples.

In the news and product and service review context, a multi-label classification method that integrated label correlations into consistency regularization was elaborated in Qiu et al. (2020). Consistency regularization contributed to the model predicting the same class for an unlabeled sample even though it was perturbed. The method leveraged the Exponential Moving Average model and the label correlation matrix to generate an accurate target for each unlabeled instance and applied the mixup technique to compute consistency regularization. Miyato et al. (2017) extended the virtual adversarial training (VAT) from images to text classification. Text embeddings suffered perturbations because VAT uses continuous inputs, approximate adversarial virtual perturbation was used which corresponded to a second-order Taylor expansion and the power method was applied.

In the product and service review context, based on the CBOW model, Zhang et al. (2020) analyzed the appropriate perturbations to generate the adversarial texts that are readable to deceive human observers by controlling the perturbation direction vectors. The perturbations meet the context in the neighborhood of words. Meanwhile, they used adversarial product and movie review texts to enhance the robustness of the model with Adversarial Training to regularize the classification model and extended it to semi-supervised tasks with VAT. The method demonstrated that the generated adversaries' texts and original texts had a similar meaning, they were interpretable and confused to humans and the VAT improved the robustness of the model. The method trained a model to defend against readable adversarial text attacks. Li and Sethy (2020) proposed a framework Layer partitioning for discrete text input which was combined Π-Model or temporal ensembling for short text classification. A neural network was split into two parts, one part with lower layers used to feature extractor and to add systematical noise in the input, and the other one with higher layers. With the perturbed input, the SSL method was used to train the higher layers employing Π-Model and temporal ensembling.

For scientific context, Sun et al. (2019b) investigated VAT to the supervised loss of GCN to improve the performance in scientific articles classification. Thus, GCN Sparse VAT (GCNSVAT) and GCN Dense VAT (GCNDVAT) algorithms were results where virtual adversarial perturbations were inserted on sparse and dense features. Also in the context of scientific articles, due to susceptibility from GCN to the perturbations, Hu et al. (2021) used Adversarial Training considering graph structure to decrease the feature perturbations impact from a neighbor node.

For the Chinese language, considering the smoothness assumption, a semi-supervised multi-class short text classifier to detect and classify emergency events with a deep learning architecture was proposed by Liu et al. (2021). Kullback–Leibler divergence measured the distance between two predictions: clean samples and their perturbed version. In Huang et al. (2020a), it was elaborated two-stage SSL framework for Chinese patent classification

based on the theory of Inventive Problem-Solving. The method used a standard LSTM, and pooling layer with soft attention and k-Max pooling for feature extraction. The method pre-trained the model with unlabeled data, then it used a mixed objective function to train the text classification model. The mixed objective function was a combination of cross-entropy, entropy minimization, and adversarial and virtual adversarial loss functions.

4.4.2 Manifolds

Manifolds are part of intrinsically semi-supervised methods. The manifold assumption says that the data points are located in the multiple lower-dimensional manifolds which comprise the input space and data points have the same label if they are located in the same lower-dimensional manifold (van Engelen and Hoos 2019).

For sentiment analysis, Gupta et al. (2018) employed learning feature representations with Doc2Vec, pre-training, and manifold regularization to train a sentiment classification model. The manifold regularization used a mix of external and in-domain data and it was applied to train a statistical model to use the labeled and unlabeled data resources. Park et al. (2019) proposed a semi-supervised distributed representation method that reflected the difference of document distributions depending on the sentiments using partially labeled documents. A new objective function obtained document embedding best suited to sentiment information for sentiment classification. Document embeddings were acquired with one restriction related to manifold assumption, and another one related to the smoothness assumption of the sentiment classifier in learned representations.

4.4.3 Generative models

Even Manifolds, Perturbation-based methods, and Generative Models are intrinsically semi-supervised. However, different from these methods, whose only objective is to deduce a function to classify data points, generative methods have the primary objective to model the process that generated the data. Mixture models, GANs, and variational autoencoders (VAE) are examples of generative model methods.

Supposing each observation from the dataset comes from one specific distribution, i.e Gaussian distribution. The maximum likelihood or EM is used to infer the parameter of distribution, such as mean and variance. Then, with the mixture generative model method the distribution $p(x, y)$ is modeled and samples can be drawn and the model can be used for classification. GANs are deep learning architectures to train generative models. GANs approach the learning of distribution with loss function based on the zero-sum game between two players (Generator and Discriminator), where the sum of player costs is zero. The Generator is trained to deceive the Discriminator with the production of samples similar to the training data distribution, while the Discriminator in a supervised way classifies the samples as reals or fakes (Goodfellow 2017).

VAE is formed by the encoder and decoder, it is a deep generative model which can generate samples using the latent space. Each data point x is treated as being generated from a vector of latent variables z. VAEs limit $p(z)$ to a simple distribution to facilitate sampling, i.e. standard multivariate Gaussian distribution. Based on data point x, the encoder establishes the parameters of $p(z \mid x)$ distribution. While the decoder performs the transformation from $p(z)$ to a more complex distribution $p(x \mid z)$. To generate reconstructions of x, a sample is drawn from the distribution, $p(z)$, thus a sample z vector is passed through the

decoder and is multiplied by the weights, added a bias, and applied an activation function. A combined cost function with Kullback–Leibler divergence between the posterior distribution $p(z \mid x)$ and some simple prior distribution $p(z)$, and the reconstruction cost of the output of the autoencoder for input data are minimized by encoder and decoder which are trained simultaneously. The decoder is used as a generative model (van Engelen and Hoos 2019).

In the news context, the generative process for both words and response variables was employed by Soleimani and Miller (2016a). The approach was a mixture of class-conditioned topic models to discover topics and predict class labels in a semi-supervised fashion based on the assumption that documents from the same class have similar topic proportions. Manifold and cluster assumption was introduced by Xie et al. (2019) to regularize the classifier in deep generative models. The methods encouraged classifier invariance to local perturbations in the data sub-manifold of each cluster and distinct classification outputs for data points in different clusters producing a discriminative ability of the classifier. Data augmentation methods through a Generator and a Filter for topic classification and sentiment analysis were proposed by Queiroz Abonizio and Barbon Junior (2020). The Generator synthesized new samples and the Filter captured high-quality ones.

Still in the news context. BERT with semi-supervised GANs were combined in Croce et al. (2020) to text classification. The Generator produced fake samples based on the data distribution and the BERT model was used as a discriminator. By leveraging the information from hierarchy labels to generate the topics, Agarwal (2021) implemented a semi-supervised hierarchical LDA: a probabilistic graphic model to discover latent topics from the news documents by Gibbs sampler. In textual anomaly detection, Steyn and de Waal (2016) enhanced the Multinomial Naive Bayes classifier with an augmented EM algorithm. For hierarchical text classification based on a generative model, Xiao et al. (2019) proposed a path cost-sensitive learning algorithm. The approach applied the EM and local maxima were obtained based on the parameters of the Naive Bayes classifier in labeled data.

For the short text classification task, using a Kernel-based Deep Architecture combined with semi-supervised GAN, Croce et al. (2019) investigated how to improve the robustness of deep architectures by exploiting an expressive space that encodes rich linguistic information. Najari et al. (2022) customized the GAN for text-based social bot detection wherein the GAN used a common LSTM layer as a shared channel between the generator and the classifier to handle the convergence limitation of traditional Seq-GAN. Spam detection based on GANs was addressed in Stanton and Irissappane (2019), the features were learned by ANN, and the method generated similar spam/non-spam reviews in relation to the training set. Multi-layer RNN with gated recurrent units was the base cell to represent the generator and the discriminator. Aghakhani et al. (2018) modified GAN for detecting deceptive reviews by means of two discriminator models and one generative model to avoid mod collapse issues by learning from both distributions of truthful and deceptive reviews. Regularized GAN (ScoreGAN) was developed in Shehnepoor et al. (2022) for fraud review detection due to the limitation of GANs with the task. The text representation was by GLoVe concatenated with a score, and the discriminator was trained to label the reviews coming from the generator.

In the context of languages other than English, Song et al. (2016) proposed a new text categorization using the Chinese language, an algorithm based on deep learning structure and semi-supervised DBN. DBN is based on Restricted Boltzmann Machines which is ANN trained in an unsupervised way with fast learning algorithm called contrastive divergence. In the fine-tuning stage, the softmax regression classifier received the output data of DBN and used the backpropagation algorithm to construct an optimized network. Liu et al.

(2020) developed a cross-domain patent retrieval with functional, technical, and domain properties. The approach applied the Chinese word segmentation tool due the fact of the particularities of the language. Naive Bayes was used as a classifier and trained according to the primary level of functional basis, and the EM algorithm as the final classifier. The automatic Chinese patent classification method was proposed in Li et al. (2017), it was based on the functional basis and Naive Bayes theory with the aim of effectively extracting the hidden information from the patent texts and to further providing this information to support the product innovation design process.

For sentiment analysis task, Duan et al. (2020) proposed a method for sentiment classification in stock message comments. The method considered the train and test set together to avoid the affection of short messages, the inferred features were more comprehensive opposing the features of traditional learning methods which only used the train set. The generative emotion model was employed and defined a text as a probability distribution over the seven-dimensional emotion space and represented the emotion as a probability distribution over words. Semi-supervised aspect-level sentiment classification based on VAE with aspect information in the encoder/decoder and aspect-level emotion classifier was proposed by Fu et al. (2019). The method only considered the aspect-category level task and Topic Word Embedding model learned aspect-specific word embedding. The method was supported by attention-based LSTM with aspect embedding as feature representation and classifier. Besides, a conditional LSTM as the decoder of VAE to introduce the text label into the decoder was applied. Sentiment classification based on conditional VAE along with attention mechanism was elaborated by Yu et al. (2019). The latent semantic information of the but-clause was integrated with the model by the integration of the attention mechanism into conditional VAE for classification improvement.

In the scientific context, for multi-label learning problems in attributed graphs to scientific document classification, Akujuobi et al. (2018) proposed a deep generative model; based on GANs, Anokye and Kahanda (2021) developed a novel method called BioSGAN for the protein-phenotype co-mention classification task; for improving the performance of AUC-optimized classifiers with scientific texts, Fujino and Ueda (2016) applied generative models to assist the incorporation of unlabeled samples in the model; for document and sentence-level class inferences, Soleimani and Miller (2016b) investigated a multi-label topic model. The method found the topics present in the corpus, learned the association between topics and class labels, labels were predicted for new documents, and performed label associations for each sentence in the documents.

4.5 Transfer learning

Considering that domain adaptation, a method of transfer learning can be divided into unsupervised and semi-supervised approaches by the availability of labeled target data (Abdi and Hasehmi 2021). In this survey, we define transfer learning as a semi-supervised approach when a method used a small amount of labeled target data and a large and sufficient unlabeled target data.

In the news context, for binary logistic regression, Wang et al. (2019) applied multiple-source deferentially private hypothesis transfer learning method. The scarce labeled target data were treated using unlabeled data with a rigorous differential privacy guarantee. The weight assigned to each source hypothesis was determined by its relationship with the target, then the negative transfer was attenuated. Li and Dai (2018) overcome the problem of small amounts labeled in target to form a validation set extracting samples from the

source dataset based on dynamic dataset regrouping. A new inductive knowledge transfer learning algorithm integrated with a modified Rank-based Reduce Error ensemble selection approach to address the different distributions in both source and target domains were used for news text classification.

In the cross-lingual task, Moon and Carbonell (2016) sought to learn new target tasks with limited label information by leveraging source datasets with heterogeneous features and label spaces. The approach mapped heterogeneous source and target labels into the same Skip-gram word embedding to obtain their semantic class relation. In cross-lingual text categorization, Huang et al. (2020b) elaborated a novel algorithm denominated heterogeneous discriminative features learning and LP to learn discriminative features with label consistency through two domain-specific projections, and LP through exploiting structural information of data.

Still in the cross-lingual task. For heterogeneous transfer learning, Sukhija and Krishnan (2019) employed a new approach, i.e. Web-induced Heterogeneous Transfer Learning with sample selection to multilingual text classification. A novel Feature Space Remapping algorithm associated the domains with heterogeneous feature and label spaces without relying on an instance or feature correspondences between the source and target domain. Based on web-induced knowledge, labels across two domains were semantically aligned, then reached the correspondence for aligning the heterogeneous features of the source and target domain. By a novel semi-supervised discriminative transfer Learning method, Kang et al. (2019) tackled the cross-language text classification. The unlabeled data in the source and target language were used to adjust the different distribution of the features in the target labeled data. In addition to a monolingual classification for an efficient transition, where the classifier was trained with labeled data in the source language.

In the sentiment analysis task, Mathapati et al. (2019) experimented with a semi-supervised method for dual sentiment analysis to the polarity shift problem associated with an adaptive domain that conducted training with scarce labeled adapted in different domains. The approach applied collaborative deep learning due to the problem of dependency between distant terms in reviews: LSTM addressed sequence prediction and CNN extracted features. For the sentiment analysis, Abdi and Hasehmi (2021) learned a new discriminative representation of the data by innovative domain adaptation technique. The instances of the source and target domains were embedded into a new feature space, thus with the samples in a common latent feature space, the method minimized the discrepancy between the source and target distribution while the structural information of the data was preserved.

Domain adaptable lexicon to sentiment analysis using maximum entropy with bipartite clustering was built by Deshmukh and Tripathy (2017). Source and target preprocessed datasets were taken as input, an adapted entropy classifier was applied, and a bipartite graph clustering algorithm between common and uncommon words was constructed. Clustering handled the mismatch between domain-specific words of the source and target domain. In multiple domains with specialized multiple sources transfer learning based on multi-instance learning, Song and Park (2018) identified intention posts. The method used positive instances to transfer the knowledge across domains, thus false negatives that affect multi-instance learning were treated.

4.6 Others

In this subsection, we describe semi-supervised methods that do not comply with the taxonomy proposed by van Engelen and Hoos (2019).

The paragraph describes the articles in which methods were applied in the news context. TSVM algorithm based on Ant Colony Optimization to solve the transduction inference SVMs optimization problem was proposed by Yu et al. (2016). Based on PUL, Sakai et al. (2017) applied area under the curve (AUC) optimization method. Unlabeled data contributed to improving the generalization performance in PU and semi-supervised AUC optimization methods without the restrictive distributional assumptions. Cheeks et al. (2016) developed a process of discovering communication frames found in online news articles with relevant socio-environmental issue contexts. NMF was combined with TF–IDF for discovering frames through the process of revealing latent relationships in articles. Customer disputes automatically according to their root causes were classified in Severin et al. (2019). Categories and their Keywords were defined in a supervised step of the method, then the disputes were placed into the appropriate categories. Thus, reducing manual labeling of a training dataset.

In the Chinese news context, a small part of documents was automatically labeled with high accuracy based on the lexical databases as external semantic resources (Xu et al. 2017). Labeled and a lot of unlabeled documents were combined to form the training data and a TSVM and Deterministic Annealing to build the SSL approach.

5 Results analysis per datasets

A comparison among machine learning methods does not produce a reliable answer due to the fact of there are several parameters involved in the learning process. In the semi-supervised method, for example, the amount of labeled and unlabeled data, evaluation metrics, and subsets of the datasets used in the experiments not always were equal. Absolutely, we do not have the pretension to judge the semi-supervised methods, otherwise, our goal is to shed some light on the area through observation. The following subsections demonstrate the semi-supervised approaches per dataset and the results achieved by the article authors. Section 5.1 presents the 20 Newsgroups dataset. Section 5.2 presents the Reuters 21578 dataset. Section 5.3 presents the Reuters RCV1 and RCV2 datasets. Section 5.4 presents the movie review datasets. Section 5.5 presents the Twitter datasets. Section 5.6 presents the Amazon, Yelp, and TripAdvisor datasets. Section 5.7 presents the scientific datasets. Section 5.8 presents the medical datasets. Section 5.9 presents the AG News, DBpedia, and WebKB datasets. Section 5.10 presents the TREC datasets. Section 5.11 presents the Chinese and Vietnamese datasets.

5.1 20 Newsgroups dataset

Results of experiments on 20 Newsgroups dataset are shown in Table 3 which has 24 articles, five of which performed experiments with ANN. SSL approaches in addition to ANN were researched by Zhao et al. (2022) that along with GCN outperformed state-of-the-art models across five benchmark datasets.

Table 3 20 Newsgroups dataset by SSL approach

Approach	Articles	Task	Labeled	Accuracy	Precision	F_1	Alternative measures
Graph-based	Yadav et al.	Multiclass	25%, 50%	0.599 0.690	–	–	–
	Widmann and Verberne	Multiclass	N = 100 N = 350	–	–	0.620 0.680	–
	Zhao et al.	Multiclass	–	0.870	–	–	–
Feature Extraction	Jiang et al.	Multiclass	27.27%	0.826	–	–	–
Cluster-then-label	Jedrzejowicz and Zakrzewska	multiclass	0.1%, 0.5%	0.762 0.751	–	–	–
	Barman and Chowdhury	Multiclass[1]	10%	0.867	–	–	–
	Guru et al.	Multiclass[2]	20%	–	–	≈ 0.890	–
	Vilhagra et al.	Multiclass	n = 1.0[3]	–	–	–	0.119 Clustering F-Score
	Sun et al.	Binary	–	0.940	–	–	–
Self-training	Pavlinek and Podgorelec	Multiclass	0.1% , 0.5%	0.734 0.733	–	–	–
	Altnel et al.	Multiclass[4]	1%, 5%, 30%	0.597 0.860 0.950	–	–	–
	Altnel et al.	Multiclass[5]	1%, 5%, 30%	0.526 0.887 0.953	–	–	–
	Altnel et al.	Multiclass[6]	1%, 5%, 30%	0.325 0.498 0.729	–	–	–
	Altınel and Ganiz	Multiclass[4]	1%, 5%, 30%	0.566 0.859 0.953	–	–	–
	Altınel and Ganiz	Multiclass[5]	1%, 5%, 30%	0.519 0.845 0.942	–	–	–
	Altınel and Ganiz	Multiclass[6]	1%, 5%, 30%	0.431 0.617 0.742	–	–	–
Co-training	Iglesias et al.	Multiclass	50%	–	–	0.710	–
	Zhang et al.	Multiclass	0.4%	–	–	–	0.937 Weighted-F1
Boosting	Liu et al.	Binary	1%[7]	–	–	≈ 0.900	–
	Liu et al.	Multiclass	1%[7]	–	–	–	≈ 0.700 Macro-avg.-F_1

Table 3 (continued)

Approach	Articles	Task	Labeled	Accuracy	Precision	F_1	Alternative measures
Generative Models	Fujino and Ueda	Binary	2%, 4%	–	–	–	0.916 Avg. AUC 0.943 avg. AUC
	Soleimani and Miller	Multiclass	10%, 30%, 50%	$\approx 0.700 \approx 0.720 \approx 0.740$	–	–	–
	Steyn and de Waal	Multiclass	30%	0.750	–	–	–
	Xiao et al.	Multiclass	1%	–	–	–	0.707 Micro-F_1, 0.600 Macro-F_1
	Croce et al.	Multiclass	1% 5% 30%	–	–	$\approx (0.45, 0.77, 0.80)$	–
Transfer Learning	Wang et al.	Binary	5%[8]	0.853	–	–	–
	Li and Dai	Binary[9]	1.8%[8]	0.971	-	-	–
Others	Yu et al.	Binary	–	–	0.915	–	–
	Sakai et al.	binary	1%	–	-	–	0.798 Avg. AUC

[1]20 Newsgroups subsets (overall acc.)

[2]D2, D9, D10 datasets

[3]Pairwise constraint

[4]Science

[5]Politics

[6]Mini-NewsGroup

[7]Per class

[8]Labeled target samples

[9](C-T) dataset

In Jiang et al. (2018), DBN surpassed the classical baseline algorithm on different data scales of datasets used beyond 20 Newsgroups. In fine-tuning optimization, L-BFGS was more adequate than gradient descent. In Vilhagra et al. (2020), CSN for the deep neural representation of the input data based on pairwise constraints outperformed the MPC-KMeans, and ordinary K-Means algorithm in six datasets, and its performance increased with the number of constraints provided. LDA and Word2Vec overcome baselines in Jedrzejowicz and Zakrzewska (2020). GAN–BERT developed by Croce et al. (2020) compared to BERT demonstrated superior results. With 1% of labeled data, GAN–BERT achieved F_1-Score higher than 40% while BERT result was below 20%. Besides, GAN–BERT was superior to baseline until 40% of labeled data.

The remaining 19 authors used algorithms other than ANN in the text representation as well as in the classification model development. In Widmann and Verberne (2017), the results were not able to prove the advantage of graph-based SSL over the supervised learning baseline. Guru et al. (2016) demonstrated the efficacy and robustness of the proposed model in detecting unknown classes efficiently. Sun et al. (2020) had superior classification accuracy over state-of-the-art SSL algorithms. Pavlinek and Podgorelec (2017) demonstrated that the self-training and LDA method when used in combination with Multinomial Naive Bayes performed the accuracy than the comparable methods. Altnel et al. (2017) labeled unlabeled instances based on meaning scores of words to augment the training set, it was valuable and increased the accuracy of previously unseen test instances. Altınel and Ganiz (2016) utilized abundant sources of unlabeled instances to improve the accuracy, especially when the number of labeled instances was limited. Iglesias et al. (2016) improved the accuracy of the text classifiers. Zhang et al. (2021a) with comparative experiments results demonstrated that the method had good classification performance. Yadav et al. (2019) compared sqrt-cosine similarity metric to Euclidean L2 norm and cosine similarity demonstrating superior results in the quality of graph construction, and the classification/inference. Barman and Chowdhury (2018) showed the effectiveness in assigning labels to a set of large unlabeled data with the help of a very small labeled dataset.

In Liu et al. (2016) the Universum supported the classifiers when few labels are available. Fujino and Ueda (2016) outperformed the baseline methods, the approach improved the imbalanced binary classification performance. Soleimani and Miller (2016a) surpassed the performance of both standard semi-supervised and supervised topic models. Steyn and de Waal (2016) had good performance with text classification. However, the results in the identification of anomalous text documents demonstrated a decreased accuracy due to the fact that unlabeled data increased the magnitude of class imbalance through EM. Xiao et al. (2019) demonstrated improvements in the algorithm's effectiveness. Wang et al. (2019) had improvement over baselines, Li and Dai (2018) outperformed the baselines non-transfer algorithms, the state-of-the-art transfer learning algorithms with lower storage requirements and higher classification speed. Yu et al. (2016) overcome the baselines of TSVM algorithms considering classification precision and running efficiency indexes. Sakai et al. (2017) exceeded with short computation time baseline algorithms.

5.2 Reuters 21578 dataset

The results with the Reuters 21578 dataset, which is a collection of documents with new articles, are presented according to Table 4. ANN was applied by four authors, three already described previously. Kumar et al. (2021) along with MLP achieved competitive

Table 4 Reuters 21578 dataset by SSL approach

Approach	Articles	Task	Labeled	Accuracy	F_1	Alternative measures
Graph-based	Rossi et al.	Multiclass	N = 10[1]	–	–	0.665[2], 0.632[3], 0.816[4] Macro-F_1
	Rossi et al.	Multiclass	N = 50[1]	–	–	0.666[2], 0.652[3], 0.885[4] Macro-F_1
	Carnevali et al.	Multiclass[2]	(N = 1 N = 5 N = 20)[1]	–	≈ (0.310 0.400 0.500)	-
	Carnevali et al.	Multiclass[3]	(N = 1 N = 5 N=20)[1]	–	≈ (0.400 0.550 0.700)	-
	Zhao et al.	Multiclass	–	0.979[4]	–	–
	Zhao et al.	Multiclass	–	0.945[5]	–	–
Feature Extraction	Jiang et al.	Multiclass	50.0%[6]	0.825[7]	–	-
	Jiang et al.	Multiclass	13.33%[6]	0.869[8]	–	–
Cluster-then-label	Jedrzejowicz and Zakrzewska	Multiclass[4]	0.1% 0.5%	0.822, 0.832	–	–
	Jedrzejowicz and Zakrzewska	Multiclass[5]	0.5%	0.500	–	–
	Thomas and Resmipriya	Multiclass	17.5%	–	–	0.890 Cosine similarity
	Thomas and Resmipriya	Multiclass	17.5%	–	–	0.930 SMTP similarity
Self-training	Villatoro-Tello et al.	Multiclass[4]	k = 1 (i = 0, i = 3)[9]	–	–	0.430, 0.650 Macro-avg.-F_1
	Villatoro-Tello et al.	Multiclass[4]	k = 5 (i = 0, i = 3)[9]	–	–	0.430, 0.730 Macro-avg.-F_1
	Kumar et al.	Binary[2]	N = 5[1]	0.430	–	–
	Kumar et al.	Binary[2]	N = 10[1]	0.590	–	–
	Pavlinek and Podgorelec	Multiclass[4]	0.1%, 0.5%	0.864, 0.850	–	–
	Pavlinek and Podgorelec	Multiclass[5]	0.5%	0.477	–	–
Co-training	Iglesias et al.	Multiclass	50%	–	0.869	–
Boosting	Tanha	Multiclass	10%	0.680[2], 0.724%[3]	–	–
	Liu et al.	Multiclass	1%	–	–	0.616 Macro-avg.-F_1
	Tanha	Multiclass	10%	0.702[2], 0.726[3]	–	–
Transfer Learning	Li and Dai	Multiclass[11]	2%[10]	0.813	–	–

Table 4 (continued)

[1]Per class

[2]Re0 [3]Re1 [4]Re8 [5]R52

[6]Labeled in fine-tuning stage

[7]Tiny Reuters

[8]Large-scale Reuters

[9] k is the number of labeled documents per iteration, and i is the iteration

[10]Labeled target samples

[11](O-Pe) dataset

performance gain in classifiers based on SSL—Cascading (gain of 7%); Rank-based (gain of 5%) over SSL baseline.

The remaining 10 authors used algorithms other than ANN in the text representation as well as in the classification model development. Four articles already had the results summarized previously. Carnevali et al. (2021) outperformed state-of-the-art algorithms based on the vector space model or graphs algorithms in terms of F_1-Score. The method improved the classification performance from 10% when using only 1 labeled document to 28% with 30 labeled documents. Rossi et al. (2017) facilitated the graph construction, Villatoro-Tello et al. (2016) demonstrated that selecting confidently labeled documents improved the performance across iterations when short text summaries were used as the set of labeled data. In Tanha (2019), Decision Tree as base learner outperformed supervised and semi-supervised baseline algorithms. Tanha (2018) surpassed state-of-the-art boosting methods to multiclass SSL. Thomas and Resmipriya (2016) had better accuracy with SMTP for the distance calculation.

5.3 Reuters RCV1 and RCV2 datasets

Reuters RCV1 and RCV2 datasets are a collection of news articles used for cross-lingual and multi-label classification. The SSL approaches results are present according to Table 5. Five authors employed the ANN approach, Li et al. (2018) used CNN for multi-label classification and improved the performance compared with traditional ANN. Shayegh et al. (2019) applied CNN and achieved results equated with several state-of-the-art supervised and SSL algorithms. In Qiu et al. (2020), pre-trained 300-dimensional fastText language model and CNN as the multi-label text classifier outperformed two supervised multi-label learning solutions, and compared with two SSL methods based on consistency regularization, the approach overcome them in 19 and 16 evaluation indicators separately. Miyato et al. (2017) with LSTM, and bi-LSTM achieved state-of-the-art performance in the RCV1 dataset with a 5.54% error rate. Besides, the method achieved state-of-the-art in various text classification tasks. Moon and Carbonell (2016) improved hetero-lingual text classification task.

The remaining seven articles used algorithms other than ANN and the results are summarized in sequence. Gong et al. (2017) overcome baselines methods in accuracy metric. Besides, the method outperformed the GFHF baseline method when label noise was present. Xu et al. (2016) with CoL(2-layer) (71.73%) and CoL(3-layer) (72.45%) outperformed the existing SSL methods which the best result achieved (69.34%). Sukhija and Krishnan (2019) outperformed the baselines SHFR-RF by 3.5–7%, SHDA-RF by 2.5–3%, DAMA by 7–15% and Co-HTL by 1.5–3.5% in every cross-lingual transfer setting. For the cross-lingual Reuters Multilingual dataset, the method had performance improvement over the baseline Random Forest, and overcome state-of-the-art transfer approaches on three diverse real-world transfer tasks. Huang et al. (2020b) outperformed several baseline adaptation methods even if the distribution difference was substantially large. Kang et al. (2019) demonstrated the overall significance of the performance with 89.2%, and 85.4% of accuracy in over 20 one-vs.-one classification tasks, and one-vs.-all classification, respectively. While the best baseline achieved 88.4%, and 84.2%, respectively.

Table 5 Reuters RCV1 and RCV2 datasets by SSL approach

Approach	Dataset	Articles	Task	Labeled	Accuracy	Precision	Alternative measures
Graph-based	RCV1	Gong et al.	Multiclass	2.5%	≈0.800	–	–
Feature Extraction	RCV1	Li et al.	Multi-label	–	–	–	0.090 Loss rate
Co-training	RCV2	Xu et al.	Multilingual	0.4%	–	0.725	–
	RCV1	Shayegh et al.	Binary	–	0.912	–	–
Perturbation-based	RCV1	Qiu et al.	Multi-label	0.3%	–	–	0.766 Micro-F_1, 0.441 Macro-F_1
	RCV1	Miyato et al.	Multiclass	–	–	–	0.067 Error rate
Generative Models	RCV1	Xiao et al.	Multiclass	1% 10% 30%	–	–	≈0.780, 0.870, 0.900 Micro-F1
	RCV1	Xiao et al.	Multiclass	1% 10% 30%	–	–	≈0.340 0.480 0.550 Macro-F_1
Transfer Learning	20[1]: RCV2	Moon and Carbonell	Cross-lingual	0.1%	0.533	–	–
	R8: RCV2	Moon and Carbonell	Cross-lingual	0.1%	0.537	–	–
	RCV2: 20[1]	Moon and Carbonell	Cross-lingual	1%	0.443	–	–
	RCV2: R8	Moon and Carbonell	Cross-lingual	0.1%	0.638	–	–
	RCV2	Sukhija and Krishnan	Cross-lingual	0.6%[2]	–	–	0.222 Avg. mean error
	RCV2	Huang et al.	Cross-lingual	0.6%[2]	0.746	–	–
	RCV2 (OvO)	Kang et al.	Cross-lingual	8.3%	0.892	–	–
	RCV2 (OvA)	Kang et al.	Cross-lingual	8.3%	0.854	–	–
Others	RCV1	Sakai et al.	Binary	1%	–	–	0.911 Avg. AUC

[1]20 Newsgroups dataset

[2]Labeled target samples

Table 6 IMDB and Movie Review (MR) datasets by SSL approach

Approach	Dataset	Articles	Task	Labeled	Accuracy	F_1	Alternative measures
Graph-based	IMDB	Ju et al.	Binary	5% 25% 50%	≈ (0.670 0.705 0.725)	–	–
	IMDB	Ju et al.	Multiclass	50%	0.433	–	–
	IMDB	Ganiz	Binary[1]	1%, 5%, 10% 30%	≈ (0.740 0.850 0.870 0.920)	–	–
	IMDB	Ganiz	Binary[2]	1%, 5% 10% 30%	≈ (0.880 0.890 0.870 0.860)	–	–
	Cornell MR	Zhao et al.	Binary	2.5% 5% 10% 50%	≈ (0.635 0.640 0.680 0.760)	–	–
	Cornell MR	Yang et al.	Binary	10% 30% 60% 80%	≈ (0.650 0.750 0.830 0.840)	–	–
Feature Extraction	IMDB	Pan et al.	Binary	0.5% 1% 4%	0.767 0.794 0.834	–	–
	IMDB	Sun et al.	Binary	0.4% 2% 4% 10%	–	–	≈ (0.092 0.080 0.055 0.053) Error rate
Cluster-then-label	IMDB subset	Vilhagra et al.	Binary	n = 0.1[3]	–	0.501	–
Self-training	IMDB	Li and Ye	Binary	N = 100 N = 500 N = 1000	0.821 0.901 0.916	–	–
	IMDB	Altnel et al.	Binary	1% 5% 30%	0.704 0.827 0.914	–	–
	IMDB	Altınel and Ganiz	Binary	1% 5% 30%	0.628 0.793 0.883	–	–
	MR	Khan et al.	Binary	–	0.861[4]	–	–
	MR	Khan et al.	Binary	–	0.857[5]	–	–
	Cornell MR	Khan et al.	Binary	–	0.850[4]	–	–
	Cornell MR	Khan et al.	Binary	–	0.865[5]	–	–
	IMDB	Xiang and Yin	Binary	N = 1250	0.890	–	–
Co-training	IMDB	Shayegh et al.	Binary	–	0.910	–	–
Perturbation-based	IMDB	Zhang et al.	Binary	25%	0.937	–	–
	IMDB	Li and Sethy	Binary	N = 100 N = 500	0.759, 0.851	–	–
	IMDB	Miyato et al.	Binary	–	–	–	0.059 Error rate
	Cornell MR	Zhang et al.	Binary	–	0.818	–	–
	Cornell MR	Miyato et al.	Binary	–	–	–	0.166 Error rate
Manifolds	MR	Gupta et al.	Binary	10% 16% 40% 56%	≈ (0.650 0.675 0.730 0.740)	–	–
Generative Model	Cornell MR[6]	Yu et al.	Binary	35%	0.807	–	–

Table 6 (continued)

[1]($\lambda = 0.5$)

[2](λ=1.0)

[3]Pairwise constraint

[4]Senti-IG

[5]Senti-Cosine

[6]Movie Review (Pang and Lee 2005) plus The Stanford Sentiment Treebank dataset

5.4 Movie review datasets

Table 6 demonstrates the results of experiments in movie review datasets, experiments with the ANN were carried out by 14 articles. Ju et al. (2022) applied GNN to learning graph representation. In IMDB multi-class dataset it was slightly lower than the baseline which had 43.7% of accuracy. In the IMDB binary dataset, varying the amounts of the labeled data, the method achieved the best performance compared to baseline algorithms. With only 5% of the labeled data, the method achieved 67.0% of accuracy roughly. GAT implemented by Yang et al. (2021a) outperformed state-of-the-art methods under both transductive and inductive learning. Pan et al. (2020) applied LN, Word2Vec, BERT, DistilBERT, or ALBERT, encoder and decoder model. The method was effective for sentiment analysis, ALBERT achieved 83.4% of accuracy considering 4% of labeled data and outperformed supervised LSTM and SVM. The cost function reduced the difference between the clean encoder and the noise encoder–decoder.

Fine-tuning pre-trained language model BERT to sentiment classification was employed by Sun et al. (2019a). The within-task and in-domain further pre-training boosted text classification performance and improved the task with small-size data. The proposed approach achieved the new state-of-the-art on eight text classification datasets. Li and Ye (2018) with the GAN approach and using neural word embeddings for text representation, LSTM as discriminator outperformed competing state-of-the-art methods. bi-GRU was implemented by Xiang and Yin (2021) and the method was compared with semi-supervised baselines demonstrating an improvement of 7% while some baselines such as Virtual Adversarial improved by 2%. However, the model achieved an accuracy of 89.0% versus 94% of accuracy from the Virtual Adversarial model. To generate the adversarial texts, Zhang et al. (2020) used CBOW and applied bi-LSTM which outperformed the methods based on adversarial training, VAT, and the baseline without perturbations. Along with the BERT language model and the ANN, Li and Sethy (2020) had results comparable to the supervised baseline.

ANN and Doc2Vec in Manifold's approach were used by Gupta et al. (2018). The method had gained in a single corpus setting as well as two cross-corpora settings, particularly when a smaller fraction of training was labeled. In two cross-corpora settings, the semi-supervised regularization outperformed baseline supervised training. With VAE and attention mechanism applied in the Generative Model approach, Yu et al. (2019) outperformed the baseline semi-supervised methods, and the method achieved an accuracy of 80.7% against the best baseline Aux-LSTM (79.5%) with 10k of unlabeled data. Aux-LSTM had better performance with 1k, 2k, and 4k of unlabeled data, but CVAE-Attention achieved the best performance with 10k of unlabeled data.

The remaining five authors from 17 articles investigated algorithms other than ANN. Ganiz (2016) with $\lambda = 1$ achieved 88.00% of accuracy in the IMDB dataset which was more than 10% difference from its closest competitor when the training dataset size was only 1.0% and unlabeled data size was 79.0%. In the 1150haber dataset, the method with $\lambda = 1$ achieved an accuracy of more than 90.0% with 1% of the data as the labeled training set. The method outperformed the baseline semi-supervised algorithm in the WebKB4 dataset, with $\lambda = 0.5$ achieving an accuracy of about 77.0% with 1.0% as the labeled training set. Khan et al. (2017) had an accuracy improvement of 2–3% on average when the model selection procedure was introduced. The approach outperformed the state-of-the-art semi-supervised and supervised approaches in the Cornell MR dataset.

Table 7 Tweeter datasets by SSL approach

Approach	Dataset	Articles	Task	Labeled	Accuracy	F_1	Alternative measures
Graph-based	Sentiment dataset	Yang et al.	Binary	10% 40% 60% 80%	≈ (0.650 0.730 0.775 0.800)	–	–
Cluster-then-label	STS[1]	Nguyen	Multiclass	10% 30% 50% 80%	0.748 0.775 0.795 0.816	–	–
	Obama–McCain Debate	Nguyen	Multiclass	10% 30% 50% 80%	0.686 0.715 0.756 0.784	–	–
	Tweeter 6 emotion	Namrutha Sridhar et al.	Multi-label	N = 3000	0.950	–	–
Feature extraction	Sanders Corpus	Baecchi et al.	Binary	32 256 768	0.663 0.763 0.830	–	–
	Sentiment140	Baecchi et al.	Binary	–	0.838	–	–
	Semeval-2013	Baecchi et al.	Binary	–	–	0.722	–
Self-training	Guns Advocacy	Stanojevic et al.	Binary	N = 5k	0.965	–	–
	Disaster Tweets (Kaggle)	Ghosh and Desarkar	Binary	–	–	–	0.770 Macro-F_1
	$FIRE_{16}$	Ghosh and Desarkar	Multiclass	3.1%	–	–	0.612 Macro-F_1
	$SMERP_{17}$	Ghosh and Desarkar	Multiclass	3.9%	–	–	0.866 Macro-F_1
	ADR	Karisani and Karisani	Binary	N = 300, N = 500	–	0.397 0.420	–
	Earthquake	Karisani and Karisani	Binary	N = 300 N = 500	–	0.737 0.752	–
	Product	Karisani and Karisani	Binary	N = 300, N = 500	–	0.766 0.787	–
	Parkinson's Disease	Hasan et al.	Multiclass	10% 25% ≥ 50%	–	–	0.684 0.815 0.875 Macro-F_1
	Parkinson's Disease	Hasan et al.	Multiclass	10% 25% ≥ 50%	–	–	0.783 0.840 0.877 Micro-F_1
Boosting	CIKM (training) and STS-Test	Hanafy et al.	Multiclass		0.861	–	–
Manifold	Sentiment140	Gupta et al.	Binary	1.0% 3.0% 30%	≈ (0.610 0.640 0.655)	–	–
Generative Models	Genuine accounts + spambots	Najari et al.	Binary	–	0.951[2]	–	–
	CyberTrolls[3 4]	Queiroz Abonizio and Barbon Junior	Binary	Orig 1.0, 3.0 9.0	≈ (0.635 0.655 0.680 0.680)	–	–

[1]Stanford Twitter Sentiment (STS)

[2]200D GloVE

[3]Original= 1.0% of labeled samples, augmentation rate (1.0, 3.0, 9.0)

[4]BERT and PREDATOR

5.5 Twitter datasets

Table 7 demonstrates the Twitter datasets in the experiments, where 9 of 12 articles applied ANN. Namrutha Sridhar et al. (2020) produced word embedding for the entire Twitter dataset by Word2Vec, and one of the based learners was MLP. The method had the best overall labels and individual class labels among the baselines. In Baecchi et al. (2015), CBOW with negative sampling and Logistic Regression improved the accuracy compared to CBOW representation. Using the fastText language model and deep learning models, Stanojevic et al. (2019) outperformed alternative algorithms by capturing additional contexts from the unlabeled data. The method was equated with state-of-the-art classification models.

In Karisani and Karisani (2021), BERT and ANN overcome the baseline algorithms in the ADR dataset, Earthquake dataset when data labeled N = 500, and in the Product dataset. The approach outperformed the existing state-of-the-art semi-supervised classifiers across multiple settings. With Word2Vec, LSTM and CNN, Hanafy et al. (2018) improved the accuracy of the individual models by more than 1% using a simple voting ensemble. The method achieved accuracy near to the state-of-the-art results with 170K of training data i.e. using only 10% of baseline models. GAN with a common LSTM implemented by Najari et al. (2022) had appropriate results for bot detection. Queiroz Abonizio and Barbon Junior (2020) used DistilGPT-2 as a generator, and DistilBERT as a discriminator to augment real-world social media datasets overcoming the recent text augmentation techniques.

The following three authors did not use ANN. Nguyen (2016) outperformed all other baseline methods by accuracy performance when only a few labeled instances were used. Ghosh and Desarkar (2020) achieved Macro-F1 of 61.18% against 58.68% from baseline, both models with SVM in FIRE16 dataset, and achieved 86.60% versus 85.23% of baseline in SMERP17 dataset. Experiments on three disaster-related datasets demonstrated that improvement results in overall performance increased over a standard supervised approach. In Hasan et al. (2020), the score was further improved for MedHelp and Twitter when symptom and side-effect classes were combined into one single class. The improvement of the Macro-F_1 and Micro-F_1 score by the semi-supervised model was about 1% when symptom and side-effect dictionaries were not used and the training size was less than 50%.

5.6 Amazon, Yelp and TripAdvisor datasets

Table 8 demonstrates the results of experiments with Amazon, Yelp, and TripAdvisor datasets by SSL approaches. From 22 articles, 16 performed experiments with the ANN at some stage of the classification task. With convolutional–deconvolutional auto-encoding, Chawla et al. (2019) outperformed the baseline for sentiment classification in the Yelp dataset with 1% of labeled data, and the state-of-the-art for text reconstruction in the Hotel review dataset as well as the Enron email data. Joint learning, with pre-training and data-relevant language, features improved the performance of the model for effect prediction in the Enron-FFP dataset. In Zaghdoudi and Glomann (2021), LSTM achieved an accuracy of about 87.0% in multi-label classification. Zhang et al. (2021b) applied BERT for the embedding in addition to classification, and GNN with attention-based aggregation. In the product categorization dataset with 683 categories and only three seed documents per category achieved accuracy which was only less than 2% from the supervised BERT model trained with about 50K labeled documents. Using Word2Vec, Park et al. (2019) had

Table 8 Amazon, Yelp, and TripAdvisor datasets by SSL approach

Approach	Dataset	Articles	Task	Labeled	Accuracy	F_1	Alternative measures
Graph-based	Yelp subset	Xu and Li	Multiclass	–	–	–	0.525 Avg. RMSE
Feature Extraction	YelpNYC[1]	Pan et al.	Multiclass	0.5% 1.0% 2.0% 4.0%	0.462 0.475 0.488 0.505	–	–
	YelpNYC[2]	Pan et al.	Multiclass	0.5% 1.0% 2.0% 4.0%	0.460 0.479 0.493 0.497	–	–
	Yelp	Chawla et al.	Binary	1.0% 10%	0.885 0.935	–	–
Self-training	Amazon[3]	Khan et al.	Binary	–	0.799[4]	–	–
	Amazon[3]	Khan et al.	Binary	–	0.805[5]	–	–
	Amazon Full	Stanojevic et al.	Multiclass	0.1% 1.4%	0.409 0.482	–	–
	Amazon Polarity	Stanojevic et al.	Binary	0.1% 1.3%	0.817 0.876	–	–
	Yelp	Stanojevic et al.	Binary	0.8% 8.4%	0.810 0.908	–	–
	Yelp Full	Li and Ye	Multiclass	N = 100 N = 500 N = 1000	0.537 0.552 0.573	–	–
	Amazon's website	Zaghdoudi and Glomann	Multilabel	2.5k	0.870	–	–
Co-training	Amazon[6]	Jing	Binary	≈ 35%	0.823	0.786	–
	Elec[7]	Shayegh et al.	Binary	–	0.905	–	–
	Amazon[8]	Zhang et al.	Binary	≈ 2%	–	–	0.921 Micro-F_1
	Amazon[8]	Zhang et al.	Binary	≈ 2%	–	–	0.905 Macro-F_1
Perturbation-based	Elec[7]	Zhang et al.	Binary	–	0.947	–	–
	Amazon-500	Qiu et al.	Multilabel	7.6%	–	–	0.46 Micro-F_1 0.18 Macro-F_1
	Elec[7]	Miyato et al.	Binary	–	–	–	0.055 Error rate
Manifolds	Amazon[9]	Park et al.	Binary	30%	0.912	–	–
	Yelp	Park et al.	Binary	30%	0.932	–	–
Generative Models	Yelp-main	Shehnepoor et al.	Binary	0.1% 0.3%	≈ 0.850 ≈ 0.849	–	–
	TripAdvisor	Aghakhani et al.	Binary	–	0.891	–	–
	TripAdvisor	Shehnepoor et al.	Binary	–	0.773	–	–
	TripAdvisor[10]	Stanton and Irissappane	Binary	10% 30% 50%	0.678 0.797 0.839	–	–

Table 8 (continued)

Approach	Dataset	Articles	Task	Labeled	Accuracy	F_1	Alternative measures
Transfer Learning	Amazon[9]	Mathapati et al.	Binary	1.0% 5.0%	0.913 0.923	–	–
	Amazon[9]	Abdi and Hasehmi	Binary	< 1%[11]	–	–	0.304 Avg. error rate
	Amazon[12]	Deshmukh and Tripathy	Binary	–	0.851	–	–
	Amazon CLS	Sukhija and Krishnan	Crosslingual	0.8%[11]	–	–	0.314 Avg. mean error
Others	Amazon2[13]	Sakai et al.	Binary	1%	–	–	0.872 Avg. AUC

[1]Using pre-trained DistlBERT

[2]Using pre-trained ALBERT

[3](Books, DVD, Health, Video)

[4]Senti-Cosine

[5]Senti-IG

[6](Household items, Computer supplies, Clothes)

[7](Elec: Amazon electronic product review)

[8](≈ 100K products from Amazon.com)

[9](Books, DVD, Electronics, Kitchen)

[10]Using spam-GAN-50% (50% unlabeled data)

[11]Target samples [12](Books, DVD, Music)

[13](Books and Music from Amazon7 dataset)

sentiment prediction better compared to traditional representations methods in both Amazon and Yelp datasets.

Using GAN, LSTM as a generator, and CNN as a discriminator, Shehnepoor et al. (2022) outperformed baseline methods. Aghakhani et al. (2018) with Word2Vec and GAN, LSTM as a generator, and CNN as a discriminator demonstrated the same performance in terms of accuracy that the state-of-the-art approaches which applied supervised machine learning. Stanton and Irissappane (2019) used word embedding generated by an ANN, and multi-layer RNN with GRUs as the base cell to represent the generator, and the RNN for the discriminator. Experiments demonstrated that the method surpassed state-of-the-art supervised and semi-supervised techniques when labeled data is limited. LSTM to address sequence prediction and CNN to extract features, Mathapati et al. (2019) demonstrated that deep collaboration had better accuracy in relation to Naive Bayes, CNN, or LSTM. Using ANN word embedding, Abdi and Hasehmi (2021) achieved superior results in comparison with unsupervised and semi-supervised state-of-the-art domain adaptation approaches.

The remained articles did not use an ANN in any stage of text classification, some of them already had the results summarized previously. Sentiment classification was improved by leveraging reviewer information, accordingly with Xu and Li (2017). In Deshmukh and Tripathy (2017), the accuracy achieved by the baseline method was 78.14% to 80.04%, whereas the accuracy of the proposed approach was from 71.65 to 96.89%.

5.7 Scientific datasets

Table 9 demonstrates the results of SSL approaches in the Scientific datasets. In the Graph-based approach and ANN, Zhu et al. (2021) applied GNN to learn different aspects of pre-trained global features and the raw attributes of the graph. The method achieved SSL state-of-the-art results in both plain and attributed graphs. With label consistency GNN, Xu et al. (2020) outperformed traditional GNNs in node classification. Wang et al. (2021) along with CNN and graph embedding branch to learn global features outperformed comparative approaches in the CiteSeer and Cora dataset with an accuracy improvement of 2.4% and 3.9%, respectively. In PubMed, the performance of the proposed model was only 0.7% lower than the baseline. Yang et al. (2021b) employed a simplified multilayer GCN where redundant computation was handled with the removal of nonlinearities and merging weight matrices between graph conventional layers. The method matched the running speed of simple graph convolution (SGC) and outperformed GCN and SGC in five downstream tasks.

Overfitting was reduced by the feature augmentation from the dropout layer by Hu et al. (2021) with CNN. Besides, the method improved the robustness effectively and generalization performance of GCNs, and it improved the performance in the scenario where rare few labels were available for training. GCNSVAT and GCNDVAT algorithms were applied by Sun et al. (2019b), and the method demonstrated the effectiveness under different training sizes across scientific datasets. Huang et al. (2021) along with GAT surpassed benchmarks and achieved the most advanced performance in Cora, CiteSeer, and PubMed. Attention network embedding by two layers of bi-GRU was applied by Liu et al. (2018a) that outperformed the baseline methods. Akujuobi et al. (2020) used a recurrent-attention strategy, the method was flexible for working in both transductive and inductive settings. In the transductive setting, the model

Table 9 Scientific datasets by SSL approach

Approach	Dataset	Articles	Task	Labeled	Accuracy	Precision	F_1	Alternative measures
Graph-based	Cora	Zhu et al.	Multiclass	(10 30 60)%	–	0.77 0.83 0.86	-	-
	CiteSeer	Zhu et al.	Multiclass	(10 30 60)%	–	0.54 0.64 0.71	–	–
	Cora	Yang et al.	Multiclass	5.2%	0.823	–	–	–
	CiteSeer	Yang et al.	Multiclass	3.6%	0.718	–	–	–
	PubMed	Yang et al.	Multiclass	0.3%	0.791	–	–	–
	Cora	Wang et al.	Multiclass	5.2%	0.651	–	–	–
	CiteSeer	Wang et al.	Multiclass	3.6%	0.671	–	–	–
	PubMed	Wang et al.	Multiclass	0.3%	0.765	–	–	–
	Cora	Huang et al.	Multiclass	5.2%	0.869	–	–	–
	CiteSeer	Huang et al.	Multiclass	3.6%	0.822	–	–	–
	PubMed	Huang et al.	Multiclass	0.3%	0.805	–	–	–
	Cora	Xu et al.	Multiclass	5.2%	0.835	–	–	–
	CiteSeer	Xu et al.	Multiclass	3.6%	0.738	–	–	–
	PubMed	Xu et al.	Multiclass	0.3%	0.791	–	–	–
	CoraL1	Akujuobi et al.	Multiclass	(10 20 50)%	0.801 0.818 0.844	–	–	–
	CoraIDA	Akujuobi et al.	Multiclass	(10 20 50)%	0.761 0.780 0.806	–	–	–
	Cora	Liu et al.	Multiclass	(10 30 50)%	–	–	–	0.852 0.871 0.873
	DBLP	Liu et al.	Multiclass	(10 30 50)%	–	–	–	0.806 0.811 0.821
	DBLP	Akujuobi et al.	Multiclass	(10 20 50)%	0.809 0.816 0.821	–	–	–
	Delve	Akujuobi et al.	Multiclass	(10 20 50)%	0.816 0.829 0.842	–	–	–
	CSTR	Carnevali et al.	Multiclass	N = (1 5 10)	–	–	≈ (0.60 0.70 0.78)	–
Feature extraction	PubMed	Agibetov et al.	Multiclass	N = (100 400)	–	–	≈ 0.620 ≈ 0.700	–
Self-training	Ohscal	Pavlinek and Podgorelec	Multiclass	0.1% 0.5%	0.658 0.667	–	–	–
Co-training	CiteSeer	Guo X	Binary	1.2%	–	–	–	VI(0.18) VII(0.30) error rate
Boosting	CiteULike	Liu et al.	inary	1.0%	0.783	–	–	–

Table 9 (continued)

Approach	Dataset	Articles	Task	Labeled	Accuracy	Precision	F_1	Alternative measures
Perturbation-based	Cora	Sun et al.	Multiclass	5.2%	0.786	–	–	–
	CiteSeer	Sun et al.	Multiclass	3.6%	0.693	–	–	–
	PubMed	Sun et al.	Multiclass	0.3%	0.776	–	–	–
	Cora	Hu et al.	Multiclass	5.2%	0.825	–	–	–
	CiteSeer	Hu et al.	Multiclass	3.6%	0.727	–	–	–
	PubMed	Hu et al.	Multiclass	0.3%	0.798	–	–	–
	AAPD[1]	Qiu et al.	Multilabel	9.0%	–	–	–	0.65 Mic-F_1 0.39 Mac-F_1
Generative Models	Cora	Fujino and Ueda	Binary	N = (100 200)	–	–	–	0.855 0.901 Avg. AUC
	PubMed[2]	Anokye and Kahanda	Binary	N = 809	–	–	–	0.917 F-Max
	Delve	Akujuobi et al.	Binary	0.6%	–	–	–	0.52 Mac-F_1 0.62 Mic-F_1

[1]ArXiv Academic Paper Dataset (AAPD)

[2]Subset labeled of ProPheno (Shahri et al. 2019) and unlabeled data from PubMed plus Medline

exhibited similar performance compared with GCN, but it outperformed all other baseline methods in all settings. Extensive experiments in four datasets demonstrated that the proposed method outperformed several state-of-the-art methods. Akujuobi et al. (2018) applied ANN and overcome the baselines. Anokye and Kahanda (2021) using MLP, and bi-LSTM achieved state-of-the-art performance for classifying the validity of a given sentence-level co-mention from biomedical literature outperforming traditional machine learning-based with an F-max of 81.0%.

5 of 18 articles used different methods. Guo (2018) achieved an error rate of about 8% after 40 iterations with View 1 and 10% after 45 iterations with View 2 in the Courses dataset, and achieved an error rate of about 9% after 30 iterations with View 1 and 5% after 30 iterations with View 2 in ads dataset. The results demonstrated that the proposed approach outperformed the original co-training and DCPE co-training in Courses and ads datasets. The remaining articles were described previously.

5.8 Medical datasets

The results from the Medical datasets in addition to SSL approaches are presented in Table 10. Two authors implemented GNN. Without ANN, Soleimani and Miller (2016b) achieved better labeling performance than baseline methods and increased the quality of topics (higher likelihood of unseen data), even compared to other semi-supervised methods such as LDA. Besides, the proposed approach outperformed several baseline methods concerning both document and sentence labeling as well as test set log-likelihood.

5.9 AG News, DBpedia, WebKB datasets

The AG News, DBpedia, and WebKB datasets results are presented in Table 11, 7 of 14 articles implemented ANN. In Xie et al. (2019), the encoder and classifier implemented were vanilla LSTM networks and the decoder applied the conditioned LSTM. Without ANN, Wu et al. (2019) performed baselines web page classification with the ratio of the number of labeled training samples to the total number of training samples increasing from 10 to 90%. Experiments with widely used web page datasets demonstrated that the proposed approach significantly outperformed state-of-the-art semi-supervised multi-view feature learning.

5.10 TREC dataset

Table 12 demonstrates the TREC dataset and SSL approaches where two of six articles used ANN. With deep neural representation, Liu et al. (2018a) outperformed the MPC-KMeans and ordinary K-Means algorithms. Along with the BERT language model and only 60 label samples, Li and Sethy (2020) had a better result than the semi-supervised ULMFiT with 100 label samples.

5.11 Chinese and Vietnamese datasets

Table 13 demonstrates the Chinese, Vietnamese datasets and SSL approaches. Ji et al. (2021) applied GNN and variants from GNN, e.g. binary sample GCN and binary sample

Table 10 Medical datasets by SSL approach

Approach	Dataset	Articles	Task	Labeled	Accuracy	F_1	Alternative measures
Graph-based	Ohsumed[1]	Yang et al.	Multiclass	–	0.427	–	–
	Ohsumed[2]	Yang et al.	Multiclass	6.2%	0.421	–	–
	Ohsumed[3]	Yang et al.	Multi-label	6.2%	–	–	0.502 Micro-F_1 0.424 Macro-F_1
	Ohsumed[4]	Yang et al.	Multi-label	6.2%	–	–	0.504 Micro-F_1 0.437 Macro-F_1
	Ohsumed[5]	Carnevali et al.	Multiclass	N = (1 5 10 20 30)	–	≈ (0.47 0.62 0.65 0.67 0.67)	–
	Ohsumed[5]	Rossi et al.	Multiclass	N = (1 5 10 20 30)	–	–	≈ (0.35 0.52 0.73 0.78 0.81) Micro-F_1
	Ohsumed[5]	Rossi et al.	Multiclass	N = (1 5 10 20 30)	–	–	≈ (0.36 0.52 0.74 0.77 0.80) Macro-F_1
	Ohsumed	Zhao et al.	Multiclass	–	0.703	–	–
Generative Models	Ohsumed	Soleimani and Miller	Multi-label	1.0% 10% 20% 40%	–	–	≈ (0.84 0.90 0.91 0.92) Micro ROC
	Ohsumed	Soleimani and Miller	Multi-label	1.0% 10% 20% 40%	–	–	≈ (0.76 0.86 0.87 0.88) Macro ROC

[1]Transductive learning

[2]Inductive learning

[3]Ohsumed-multi dataset and Transductive learning

[4]Ohsumed-multi dataset and Inductive learning

[5]F1-Score averaged from Oh0, Oh5, Oh10, Oh15 dataset

Table 11 AG News, DBpedia, WebKB datasets by SSL approach

Approach	Dataset	Articles	Task	Labeled	Accuracy	F_1	Alternative measures
Graph-based	AG New[12]	Yang et al.	Multiclass	1.3%	0.702	–	–
	AG News[23]	Yang et al.	Multiclass	–	0.721	–	–
	WebKB4	Ganiz	Multiclass	1.0% 10% 30%	≈ (0.770 0.870 0.878)	–	–
	WebKB	Rossi et al.	Multiclass	0.1% 1.0% 2.5% 4.2%	–	–	0.222 0.380 0.448 0.469 Micro-F_1
	WebKB	Rossi et al.	Multiclass	0.1% 1.0% 2.5% 4.2%	–	–	0.267 0.437 0.493 0.514 Macro-F_1
Cluster-then-label	AG News[2]	Vilhagra et al.	Multiclass	n = 1.0[4]	–	0.691	–
	DBpedia	Vilhagra et al.	Multiclass	n = 1.0[4]	–	0.691	–
	WebKB	Jedrzejowicz and Zakrzewska	Multiclass	0.1% 0.5%	0.757 0.756	–	–
	WebKB	Sun et al.	–	1[5]	≈ 0.900	–	–
Self-training	AG News	Li and Ye	Multiclass	N = 100 N = 500 N = 1000	0.817 0.915 0.926	–	–
	AG News	Stanojevic et al.	Multiclass	3.9% 39.0%	0.855, 0.884	–	–
	DBpedia	Li and Ye	Multiclass	N = 100 N = 500 N = 1000	0.985 0.988 0.990	–	–
	WebKB4	Pavlinek and Podgorelec	Multiclass	0.1% 0.5%	0.740 0.739	–	–
	WebKB	Wu et al.	Binary	10% 30% 50%	≈ (0.910 0.960 0.975)	–	–
Boosting	WebKB	Liu et al.	Multiclass	1%[6]	–	–	0.604 Macro-avg.-F_1
Generative Models	AG News[7]	Queiroz Abonizio and Barbon Junior	Multiclass	org.[8] 1.0, 3.0 9.0	≈ (0.872 0.880 0.880 0.880)	–	–
	AG News	Xie et al.	Multiclass	6.3% 12.5% 25.1%	–	–	0.080 0.074 0.065 Error rate
	AG News[2]	Soleimani and Miller	Multiclass	10% 30% 50%	≈ (0.620 0.640 0.660)	–	–
	DBpedia	Soleimani and Miller	Multiclass	10% 30% 50%	≈ (0.920 0.925 0.930)	–	–
	DBpedia	Soleimani and Miller	Multilabel	1.0% 10% 20%	–	–	≈ (0.770 0.900 0.900) Micro ROC
	DBpedia	Soleimani and Miller	Multilabel	1.0% 10% 20%	–	–	≈ (0.770 0.890 0.890) Macro ROC

Table 11 (continued)

[1]Inductive learning

[2]Subset of AG News

[3]Transductive learning (subset of AG News)

[4]Pairwise constraint

[5]Labeled datum

[6]Per class

[7]BERT and PREDATOR

[8]Original = 0.3% of labeled samples, augmentation rate (1.0, 3.0, 9.0)

Table 12 TREC dataset by SSL approach

Approach	Dataset	Articles	Task	Labeled	Accuracy	F_1	Alternative measures
Graph-based	TREC (tr11)	Carnevali et al.	Multiclass	N=(1, 10, 30)	–	≈ (0.520 0.700 0.790)	–
	TREC (tr12)	Carnevali et al.	Multiclass	N = (1, 10, 30)	–	≈ (0.550 0.690 0.810)	–
	TREC (tr21)	Carnevali et al.	Multiclass	N = (1, 10, 30)	–	≈ (0.550 0.700 0.800)	–
	TREC (tr23)	Carnevali et al.	Multiclass	N = (1, 10, 30)	–	≈ (0.500 0.580 0.600)	–
	TREC (tr31)	Carnevali et al.	Multiclass	N = (1, 10, 30)	–	≈ (0.510 0.800 0.780)	–
	TREC (tr41)	Carnevali et al.	Multiclass	N = (1, 10, 30)	–	≈ (0.550 0.810 0.790)	–
	TREC (tr45)	Carnevali et al.	Multiclass	N = (1, 10, 30)	–	≈ (0.600 0.720 0.810)	–
	TREC (tr11)	Rossi et al.	Multiclass	N = (1, 10, 30)	–	–	≈ (0.300 0.800 0.820) Micro-F_1
	TREC (tr11)	Rossi et al.	Multiclass	N = (1, 10, 30)	–	–	≈ (0.250 0.650 0.645) Macro-F_1
	TREC (tr12)	Rossi et al.	Multiclass	N = (1, 10, 30)	–	–	≈ (0.250 0.770 0.820) Micro-F_1
	TREC (tr12)	Rossi et al.	Multiclass	N = (1, 10, 30)	–	–	≈ 0.200 0.720 0.750 Macro-F_1
	TREC (tr21)	Rossi et al.	Multiclass	N = (1, 10, 30)	–	–	≈ 0.410 ≈ 0.850 ≈0.850 Micro-F_1
	TREC (tr21)	Rossi et al.	Multiclass	N = (1, 10, 30)	–	–	≈ 0.180 ≈ 0.500 ≈ 0.460 Macro-F_1
	TREC (tr23)	Rossi et al.	Multiclass	N = (1, 10, 30)	–	–	≈ 0.430 ≈ 0.820 ≈ 0.910 Micro-F_1
	TREC (tr23)	Rossi et al.	Multiclass	N = (1, 10, 30)	–	–	≈ 0.360 ≈ 0.650 ≈ 0.720 Macro-F_1
	TREC (tr31)	Rossi et al.	Multiclass	N = (1, 10, 30)	–	–	≈ 0.580 ≈ 0.920 ≈ 0.950 Micro-F_1
	TREC (tr31)	Rossi et al.	Multiclass	N = (1, 10, 30)	–	–	≈ 0.480 ≈ 0.750 ≈ 0.800 Macro-F_1
	TREC (tr41)	Rossi et al.	Multiclass	N = (1, 10, 30)	–	–	≈ 0.500 ≈ 0.890 ≈ 0.910 Micro-F_1
	TREC (tr41)	Rossi et al.	Multiclass	N = (1, 10, 30)	–	–	≈ 0.500 ≈ 0.800 ≈ 0.830 Macro-F_1
	TREC (tr45)	Rossi et al.	Multiclass	N = (1, 10, 30)	–	–	≈ 0.490 ≈ 0.890 ≈ 0.910 Micro-F_1
	TREC (tr45)	Rossi et al.	Multiclass	N = (1, 10, 30)	–	–	≈ 0.440 ≈ 0.835 ≈ 0.860 Macro-F_1
Cluster-then-label	TREC-6	Vilhagra et al.	Multiclass	n = 1.0[1]	–	0.558	–

Table 12 (continued)

Approach	Dataset	Articles	Task	Labeled	Accuracy	F_1	Alternative measures
Boosting	TREC (tr11)	Tanha	Multiclass	10%	0.796	–	–
	TREC (tr31)	Tanha	Multiclass	10%	0.931	–	–
	TREC (tr45)	Tanha	Multiclass	10%	0.892	–	–
	TREC (tr31)	Tanha	Multiclass	10%	0.901	–	–
	TREC (tr41)	Tanha	Multiclass	10%	0.885	–	–
Perturbation-based	TREC-6	Li and Sethy	Multiclass	1.0% 6.7%	0.678 0.860	–	–

[1]Pairwise constraint

Table 13 Chinese -zh, and Vietnamese -vi datasets by SSL approach

Approach	Dataset	Articles	Task	Labeled	Accuracy	F_1	Alternative measures
Graph-based	Sina Weibo (zh)	Guo et al.	Binary	40%	0.712	–	–
	Weibo[1] (zh)	Ji et al.	Binary	–	0.820	–	–
	Weibo[2] (zh)	Ji et al.	Binary	–	0.810	–	–
	Sogou (zh)	Zhang et al.	Multiclass	–	0.958	–	–
	Sina News[3] (zh)	Zhu et al.	Multiclass	N = 100	0.730	–	–
	Sina News[4] (zh)	Zhu et al.	Multiclass	N = 100	0.750	–	–
Cluster-then-label	Hotel reviews (vi)	Ha et al.	Multilabel	(10 20 40 60)%	–	0.58 0.66 0.74 0.77	–
	Hotel reviews (vi)	Ha et al.	Multilabel	N = (20 40 60 80)	–	–	(0.579 0.661 0.673 0.666) Micro-avg AUC
	Hotel reviews (vi)	Ha et al.	Multilabel	N = (20 40 60 80)	–	–	(0.564 0.621 0.599 0.619) Macro-avg AUC
Self-training	Dataset 1[5] (vi)	Nguyen Nhat Dang and Duong	Binary	N = 1	–	0.697	–
	Dataset 2[6] (vi)	Nguyen Nhat Dang and Duong	Multiclass	N = 1	–	0.530	–
	Dataset 3[7] (vi)	Nguyen Nhat Dang and Duong	Binary	N = 1	–	0.654	–
	Dataset 4[7] (vi)	Nguyen Nhat Dang and Duong	Binary	N = 1	–	0.624	–
	Dataset 1[5] (vi)	Duong and Anh	Binary	N = 100	–	0.861	–
	Dataset 2[6] (vi)	Duong and Anh	Multiclass	N = 100	–	0.740	–
	Dataset 3[7] (vi)	Duong and Anh	Binary	N = 100	–	0.824	–
	Dataset 4[7] (vi)	Duong and Anh	Binary	N = 100	–	0.818	–
	Sina Weibo[8] (zh)	Yin et al.	Binary	–	–	0.790	–
Perturbation-based	Weibo (zh)	Liu et al.	Multiclass	N = (70 140 210 350)	0.735 0.773 0.882 0.910	–	–
Generative Models	Sogou (zh)	Song et al.	Binary	–	–	0.983	–

Table 13 (continued)

Approach	Dataset	Articles	Task	Labeled	Accuracy	F_1	Alternative measures
Others	SogouC-UTF8 (zh)	Xu et al.	Binary	–	0.946	–	–

[1]Social traffic event topical classification

[2]Streaming social traffic event detection and analysis

[3]Sina News is labeled and Wikipedia is unlabeled. Sohu News as test set

[4]Sina News is labeled and Wikipedia is unlabeled. Classification corpus of Fudan University as test set

[5]Watch reviews

[6]Phone reviews

[7]Food reviews

[8]Sina Weibo plus Tencent Weibo

GAT. The proposed method was superior to most text classification methods in streaming social traffic event detection. Along with DBN, Song et al. (2016) extracted abstract features resulting in improved performance of the classifier which was better than the SVM algorithm. BERT language model and deep learning were employed by Liu et al. (2021). With 700 labeled examples BERT achieved 87.5% and the proposed approach 92.1% of accuracy on the Weibo dataset.

The remaining 12 articles did not apply ANN. Guo et al. (2016) achieved an accuracy improvement of 2.8% and an overall of 5.2% and outperformed the baselines in detecting credible influenza posts on Sina Weibo. Zhang et al. (2019a) achieved a classification accuracy of 96.7% and 98.1% in PKU, and FD datasets, respectively, and outperformed the best baseline algorithm. Using Sohu News and texts from Fudan University datasets, Zhu et al. (2018) outperformed the baseline method with 30% of sample expansion. Based on the expansion of 100 samples, WSE with Naive Bayes achieved the best result with an F-measure of 72.5% approximately in the Sohu dataset. WSE with SVM achieved the best result in text from Fudan University with F_1-Measure of 75% approximately. In Ha et al. (2018b), when the size of the current dataset was small, the improvement was about 2%. The proposed approach outperformed the baseline approach for all groups of experiments with an improvement of about 1%. The features built from the approach were the support for the classification and achieved the best result of 78.77% with 20 topics.

In Ha et al. (2018a), experiments in two datasets, Vietnamese reviews and English emails of Enron, demonstrated positive effects. Accuracies of classifiers for almost all experimented datasets were improved by Nguyen Nhat Dang and Duong (2019). With F_1-Score of 86.2% and the Easy Data Augmentation techniques, Duong and Anh (2021) improved Vietnamese sentiment polarity, the result achieved F_1-Score of 85.2%. Yin et al. (2018) achieved better results compared with the kNN and SLAS algorithm in five aspects, politics, economy, education, entertainment, and science and technology. Xu et al. (2017) achieved more than 95% of accuracy with until 10% of labeled documents. TSVM with 96.3% of accuracy and DA (96.6%) achieved the best results in the Netease Dataset 1 versus 86.8% from baseline SVM. In the Netease Dataset 2, TSVM had (95.8%), and DA (96.7%) versus 92.3% from baseline SVM. In the Sogou Dataset 1, TSVM had (92.6%), and DA (94.6%) compared with 94.7% from baseline SVM. Lastly, TSVM had (96.5%), and DA (96.4%) in Sogou Dataset 2 versus 93.2% from baseline SVM.

6 Benefits and limitations of the works

The benefits and limitations of each category of SSL approaches are described as follows.

GNN necessitates a huge amount of labeled data to learn effective graph representations to support graph similarity for prediction. Accordingly, with Xu et al. (2020), GCN is limited in aggregating the information from nodes with similar features or attributes for the reason that the aggregation matrix exclusively depends on graph structure. Despite great results with GAT, the aggregation matrix is based on exclusively neighboring nodes, and Brody et al. (2021) demonstrated that attention from GAT is limited, i.e. it is static attention.

In the cluster-then-label approach, low-dimensional and dense feature space is the appropriate mold to improve clustering algorithms since high-dimensional and sparsity in document clustering declines text classification performance. In a short text scenario considering document length, the problem is more accentuated, the features are

high-dimensional and extremely sparse. Besides, another problem is related to the side information, constraints quality are fundamental in semi-supervised clustering algorithms.

In the feature extraction approach, the embeddings are learned from text regions of the unlabeled data, and then applies a neural network to the supervised part. The unsupervised part of the feature extraction approach takes advantage of contextual or static features and integrates them into a supervised ANN. The approach has used a pre-trained language model (Word2Vec, BERT, among others) or ANNs such as CNN, and DBN using embedding layers to handle text input. However, Word2Vec and static embeddings are limited in relation to the permanence of the full meaning of documents, they do not recognize elements with the same meaning in different sentences and they do not treat polysemy. Furthermore, there is a dependence on a huge Corpus and they do not comprise words outside the vocabulary of the training Corpus, with the exception of fastText which solved the problem of unknown words using n-gram at character level (Kowsari et al. 2019). With the emergence of contextual text representation and transformer-based language models, words began to be interpreted from their contexts. However, the transformer and Attention mechanism face a problem to track long sequences, and large amounts (millions or billions) of parameters used and/or Corpus size make the training expensive and slow.

Self-training approach seeks to select samples confidently predicted to augment the training set. However, the threshold parameter is not always suitable for to sample selector, and without confidently pseudo-labels selection, errors influence the classifier to learn from noise, i.e errors re-enforce themselves. Another limitation is the scarce labeled data. In the transfer learning approach, a problem in domain adaptation is the discrepancy between labeled source and target instances, pseudo-label strategies to unlabeled target samples are a way to handle the problem. However, pseudo-labels are subjected to noisy information.

The co-training approach trains two classifiers on the same training data with different views for each classifier, based on the assumption that the training data has two independent views. The views are limited by methods of the text's representativeness, and contextual text representations to generate independent views integrated co-training approach has still been not much-investigated (Graef 2021), as well as deep networks as essential learn algorithm. Furthermore, co-training has the same problem in the self-training approach, when not confident unlabeled samples are added in the labeled training. In boosting approach, the pairwise similarity function is applied to labeled and unlabeled data and thus assigns more reliable pseudo-labels to unlabeled examples. However, inappropriate similarity measure compromises the algorithm performance (Tanha 2019).

In relation to the perturbed-based approach, continuous word embeddings are used in adversarial training for allowing infinitesimal perturbations due to the discrete nature of the text and its representation in high-dimensional one-hot vectors. Perturbations in texts are more difficult than image domain which is space continuous. The perturbation in texts affects the quality of examples in reason of the problem of non-interpretable adversarial examples. Models are trained to be smooth with examples based on adversarial direction, i.e. the direction where the model is more vulnerable. In a white-box attack, the generation of adversaries is a gradient-based method on word embeddings, then the quality of adversaries is linked with distance metrics. In VAT the perturbation generated is rigid due to random initialized perturbation and constraint problems.

In the Generative Model approach, GANs was very applied (36.36%). GANs have some issues that have not been completely resolved, e.g., text quality, mode collapse, training instability, and vanishing gradient. Partial collapse is more common than mode collapse, it occurs when the generator produces realistic and diverse samples, but the diversity is much less than real data distribution. GANs have problems with convergence, parameter updates

change the cost functions of the discriminator and generator and gradient climb may occur for one player and gradient descent for another player. For some games, the gradients converge and the equilibrium is achieved. However, according to the game, it is not always possible to reach equilibrium.

We observed that 78 (49.68%) articles were published in the context of short texts from a social network, product and service review, and forum discussion to investigate tasks such as sentiment analysis, emergence event detection, fake news detection, and question classification. A short text is too sparse and had an exiguous language structure, which makes it still a challenging problem for a deep neural network whose performance comes from the structured corpus. If the feature set construction does not fully represent the text, consequently sentiment analysis tasks are affected. Then, the high-dimensional sparseness of features from short texts can be further explored.

Another limitation observed in the area is related to the use of languages different from English. Only 23% of the works explored other languages like Chinese, Vietnamese, Italian, Portuguese, etc. This way, it can be difficult to find resources for other languages, such as pre-trained language models in different languages, corpora, etc. Related to this, is the few amounts of works exploring multi-lingual classification (around 1%). Moreover, the oriental languages are very different from the occidental, this way, the actual language models cannot be effective for these languages.

The percentage of labeled data varies a lot, from less than 1 to 50%. Even in the same dataset, there is no consensus on using a fixed percentage of labeled data which difficult the comparison the works. This also happens with the evaluation metrics, only accuracy is the most used, and precision and recall almost no paper calculate them. This is a limitation in the area since many papers explore multiclass classification and accuracy is not the indicated measure in this case.

7 Current research trends in SSL text classification

We identified six main future trends: ANN language model for text representation, algorithms for hyper-parameter optimization, explainable artificial intelligence (XAI) (set of methods that allows users to better comprehend the results and output created by machine learning algorithms), regularization method, development of resources for languages different from English, and analysis of degradation performance in SSL proportional to unlabeled samples.

Considering the techniques employed for text representation, it has been a growth of ANN models for generating word embeddings. Especially after 2019, the number of ANN papers surpassed the traditional algorithms, as shown in Fig. 10. Since Word2Vec, different models have been proposed like ELMo, BERT, AlBERT, GPT-2, GPT-3. Accordingly, to Fig. 8, Word2Vec and its extensions had grown since 2016, meanwhile, from 2019, they practically stabilized. Context-sensitive pre-trained model BERT appeared in 2019, and ELMo in 2020, totalizing 16 articles. However, experiments with word embedding as a part/layer of the deep learning model were the most applied compared to the word embedding language model.

We provide visualizations and analysis showing that the learned word embeddings have improved in quality and the model is less prone to overfitting. There has been a strong focus on ANN for text representation and is a current trend. The models can capture semantic and syntactic information in local sequences of consecutive words. However, they may not

capture global co-occurrences of words. New approaches using GNNs can overcome some of these problems and can be a new field to be explored. These models can lead to high accuracy by capturing contextual, semantic and syntactic properties from texts. However, it is needed to consider the limitation of GNNs in integrating the information from nodes with similar features because the adjacency matrix exclusively depends on the graph structure. In addition, it may lead to an inappropriate performance in sentiment analysis if we do not consider the order of words.

GNN has been used with an attention mechanism to construct an aggregation matrix based on embedding information. The method can benefit from the improvement of the language model and could comprise better the relationships among nodes in the graph structure. Nevertheless, more investigation is necessary to go beyond neighboring nodes in adjacency matrix formation. Besides, GNNs are computationally expensive for training and need large corpora.

Due to the discrete nature of textual data, perturbations are applied in continuous word embeddings generating a lack of interpretability. Thus, in this case, Adversarial Training is applied as a regularization method. VAT is an extension of Adversarial Training for semi-supervised text classification, problems with VAT were investigated by Li and Qiu (2020) and the results demonstrated improvement. However, VAT is a field that can be further investigated considering contextual perturbation in texts and the gradient-based method.

Adversarial Training, GANs, and contextual embeddings can be combined and exploited in the semi-supervised text classification domain. GANs suffer from the instability problem, and research has required efforts to stabilize GANs, among them are GANs and Adversarial Training associated to improve the robustness of the discriminator and training stabilization of GANs applied in image datasets (Sajeeda and Hossain 2022). However, we did not find the mixed methods in a semi-supervised text classification domain. Furthermore, pre-trained language models of domain-specific could bring improvement over the general domain. BERT, ELECTRA, and GPT families of a general and specific domain could be investigated along with Adversarial Training and GANs.

Few articles investigated the problem of degradation performance in relation to unlabeled data. Self-training suffers from semantic drift problem, Karisani and Karisani (2021) used two-stage training to cope with this problem and showed that while the number of unlabeled samples grew the performance did not drop. Altınel and Ganiz (2016) and Altınel et al. (2017) took advantage of unlabeled samples, however, the analysis in relation to growing the labeled samples and decreasing the unlabeled samples in various datasets showed better performance. Using GANs for opinion spam detection, Stanton and Irissappane (2019) demonstrated a slightly decreased in performance when the number of unlabeled samples increased. However, there is still space to investigate the potential performance degradation when considering the unlabeled data.

Other subjects that also need more investigation in semi-supervised text classification are algorithms for hyper-parameter optimization and XAI. We realized a gap in the studies regarding automated hyper-parameter tuning, and interpretability to recognize the behavior of the models. There has been an increased interest in explainability in some domains, such as medical diagnosis or legal areas. Although exist some models for explainable machine learning for models trained in text, we find few works exploring a conceptual understanding of embedding generation and the SSL models or exploring explainable IA for text classification.

Additionally, the long road ahead demands the exploration of new languages and the development of resources for languages different from English. Interdisciplinary research approaches involving applications in multiple fields probably will increase too.

8 Conclusions

Semi-supervised text classification is gaining pro-eminence due to its ability to reduce annotation costs and achieve competitive results. This survey filled the gap on this topic by selecting 157 articles from 2017 to 2022. We presented the main classification algorithms and results, datasets, SSL approaches, as well their limitations.

This study only focuses on techniques based on SSL for text classification and did not address supervised and unsupervised approaches. From the papers retrieved, it is impractical to indicate a specific classifier for a particular problem. However, various text classification techniques have been identified in different applications and the information provided in this study can help to guide the choice of the best approaches to be considered.

This survey also helps to diffuse the datasets used in the area of SSL text mining and presents in Tables 3-13 all the datasets cited in the papers, besides some information related to the approach and results obtained by the works. Especially in Table 13, we present datasets in languages other than English, to incentive more researchers to use them.

Finally, we also present many research trends that can be taken into consideration by researchers and professionals in the area.

Author Contributions JMD: Conceptualization, Methodology, Data curation, Writing—original draft. LB: Conceptualization, Writing—original draft, Writing—review and editing, Supervision.

Funding This work has been supported by the CAPES and CNPq Brazilian Research Agencies.

Declarations

Competing interests We declare that this work does not have competing interests.

References

Abdali S, Shah N, Papalexakis E (2021) Semi-supervised multi-aspect detection of misinformation using hierarchical joint decomposition. In: Machine learning and knowledge discovery in databases. Applied data science and demo track. ECML PKDD 2020, pp 406–422. ISBN 978-3-030-67669-8

Abdi L, Hasehmi S (2021) Binary domain adaptation with independence maximization. Int J Mach Learn Cybern 12:09

Abonizio QH, Junior BS (2020) Pre-trained data augmentation for text classification. In: Intelligent systems, 2020. Springer, pp 551–565. ISBN 978-3-030-61377-8

Agarwal R (2021) Phrases based document classification from semi supervised hierarchical LDA. In: 2021 2nd International conference on computation, automation and knowledge management (ICCAKM), 2021, pp 332–337

Aghakhani H, Machiry A, Nilizadeh S, Kruegel C, Vigna G (2018) Detecting deceptive reviews using generative adversarial networks. In: 2018 IEEE security and privacy workshops (SPW), 2018, pp 89–95

Agibetov A, Blagec K, Xu H, Samwald M (2018) Fast and scalable neural embedding models for biomedical sentence classification. BMC Bioinform 19:541

Akujuobi U, Sun K, Zhang X (2018) Mining top-k popular datasets via a deep generative model. In: 2018 IEEE international conference on big data (Big Data), 2018, pp 584–593

Akujuobi U, Zhang Q, Yufei H, Zhang X (2020) Recurrent attention walk for semi-supervised classification. In: Proceedings of the 13th international conference on web search and data mining, WSDM 20, 2020, pp 16–24. ISBN 9781450368223

Alam F, Joty S, Imran M (2018) Graph based semi-supervised learning with convolution neural networks to classify crisis related tweets. In: Twelfth international AAAI conference on web and social media, 2018

Alnashwan R, Sorensen H, O'Riordan A (2019) Classification of online medical discourse by modified co-training. In: 2019 IEEE fifth international conference on big data computing service and applications (BigDataService), 2019, pp 131–137
Altınel B, Ganiz M (2016) A new hybrid semi-supervised algorithm for text classification with class-based semantics. Knowl-Based Syst 108:06
Altnel B, Ganiz MC, Diri B (2017) Instance labeling in semi-supervised learning with meaning values of words. Eng Appl Artif Intell 62(C):152–163. ISSN 0952-1976
Anokye F, Kahanda I (2021) BioSGAN: protein-phenotype co-mention classification using semi-supervised generative adversarial networks. In: 2021 IEEE 34th international symposium on computer-based medical systems (CBMS), 2021, pp 468–473
Baecchi C, Uricchio T, Bertini M, Del Bimbo A (2015) A multimodal feature learning approach for sentiment analysis of social network multimedia. Multimed Tools Appl 75:05
Banerjee D, Prabhat G, Bhowal R (2018) iCASSTLE: imbalanced classification algorithm for semi supervised text learning. In: 2018 17th IEEE international conference on machine learning and applications (ICMLA), 2018, pp 1012–1016
Banitalebi-Dehkordi A, Gujjar P, Zhang Y (2022) AuxMix: semi-supervised learning with unconstrained unlabeled data. arxiv:2206.06959
Barman D, Chowdhury N (2018) A novel semi-supervised approach for text classification. Int J Inf Technol 12:1–11
Benamira A, Devillers B, Lesot E, Ray AK, Saadi M, Malliaros FD (2019) Semi-supervised learning and graph neural networks for fake news detection. In: International conference on advances in social networks analysis and mining, 2019. IEEE, pp 568–569
Billal B, Fonseca A, Sadat F, Lounis H (2017) Semi-supervised learning and social media text analysis towards multi-labeling categorization. In: 2017 IEEE international conference on big data (Big Data), 2017, pp 1907–1916
Bose J, Mukherjee S (2019) Semi-supervised method using Gaussian random fields for boilerplate removal in web browsers. In: 2019 IEEE 16th India Council international conference (INDICON), 2019, pp 1–4
Brody S, Alon U, Yahav E (2021) How attentive are graph attention networks? https://doi.org/10.48550/arXiv.2105.14491
Buza K, Revina A (2020) Speeding up the success approach for massive industrial datasets. In: 2020 International conference on INnovations in Intelligent SysTems and Applications (INISTA), 2020, pp 1–6
Carnevali JC, Rossi RG, Milios E, de Andrade Lopes A (2021) A graph-based approach for positive and unlabeled learning. Inf Sci 580:655–672. ISSN 0020-0255
Charalampakis B, Spathis D, Kouslis E, Kermanidis K (2016) A comparison between semi-supervised and supervised text mining techniques on detecting irony in Greek political tweets. Eng Appl Artif Intell 51:50–57. ISSN 0952-1976
Chawla K, Khosla S, Chhaya N (2019) Gated convolutional encoder–decoder for semi-supervised affect prediction. In: Advances in knowledge discovery and data mining, 2019. Springer, Cham, pp 237–250
Cheeks LH, Stepien TL, Wald DM (2016) Discovering news frames: exploring text, content, and concepts in online news sources to address water insecurity in the southwest region. In: 2016 IEEE 17th international conference on information reuse and integration (IRI), 2016, pp 454–462
Cheng Y, Song F, Qian K (2021) Missing multi-label learning with non-equilibrium based on two-level autoencoder. Appl Intell 51:6997–7015
Cozman F, Cohen I (2002) Unlabeled data can degrade classification performance of generative classifiers. Florida AI Research Society
Croce D, Castellucci G, Basili R (2019) Kernel-based generative adversarial networks for weakly supervised learning. In: AI*IA 2019—advances in artificial intelligence. AI*IA 2019. Lecture notes in computer science, 2019, vol 11946, pp 336–347. ISBN 978-3-030-35165-6
Croce D, Castellucci G, Basili R (2020) GAN-BERT: generative adversarial learning for robust text classification with a bunch of labeled examples. In: Proceedings of the 58th annual meeting of the Association for Computational Linguistics, 2020, online. Association for Computational Linguistics, pp 2114–2119
Day NE (1969) Estimating the components of a mixture of normal distributions. Biometrika 56(3):463–474
De Souza M, Nogueira B, Rossi R, Marcacini R, dos Santos B, Rezende S (2021) A network-based positive and unlabeled learning approach for fake news detection. Mach Learn 111(10):3549–3592
Dean B (2022) How many people use Twitter in 2022? (New Twitter stats). http://www-cs-faculty.stanford.edu
Deng X, Li Y, Weng J, Zhang J (2019) Feature selection for text classification: a review. Multimed Tools Appl 78(3):3797–3816

Deocadez R, Harrison R, Rodriguez D (2017) Automatically classifying requirements from App Stores: a preliminary study. In: 2017 IEEE 25th international requirements engineering conference workshops (REW), 2017, pp 367–371

Deshmukh JS, Tripathy AK (2017) Text classification using semi-supervised approach for multi domain. In: 2017 International conference on nascent technologies in engineering, 2017, pp 1–5

Di Capua M, Petrosino A (2017) A deep learning approach to deal with data uncertainty in sentiment analysis. In: Fuzzy logic and soft computing applications. WILF 2016. Lecture notes in computer science, vol 10147, pp 172–184. ISBN 978-3-319-52961-5

Duan J, Luo B, Zeng J (2020) Semi-supervised learning with generative model for sentiment classification of stock messages. Expert Syst Appl 158:113540. ISSN 0957-4174

Duarte JM, Sousa S, Milios E, Berton L (2021) Deep analysis of word sense disambiguation via semi-supervised learning and neural word representations. Inf Sci 570:278–297

Duong H-T, Nguyen A (2021) A review: preprocessing techniques and data augmentation for sentiment analysis. Comput Soc Netw 8:1

Felix N, Coletta LFS, Hruschka ER (2016) A survey and comparative study of tweet sentiment analysis via semi-supervised learning. ACM Comput Surv 49(1):1–26

Fujino A, Ueda N (2016) A semi-supervised AUC optimization method with generative models. In: 2016 IEEE 16th international conference on data mining (ICDM), 2016, pp 883–888

Fu X, Wei Y, Xu F, Wang T, Lu Y, Li J, Huang JZ (2019) Semi-supervised aspect-level sentiment classification model based on variational autoencoder. Knowl-Based Syst 171:81–92. ISSN 0950-7051

Ganiz MC (2016) Semi-supervised learning using higher-order co-occurrence paths to overcome the complexity of data representation. In: 2016 IEEE international conference on systems, man, and cybernetics (SMC), 2016, pp 002242–002247

Geraci F, Papini T (2018) Approximating multi-class text classification via automatic generation of training examples. In: Computational linguistics and intelligent text processing. Springer, Cham, pp 585–601. ISBN 978-3-319-77116-8

Ghosh S, Desarkar MS (2020) Semi-supervised granular classification framework for resource constrained short-texts: towards retrieving situational information during disaster events. In: 12th ACM conference on web science, WebSci '20, 2020, pp 29–38. ISBN 9781450379892

Gokhale R, Fasli M (2017) Deploying a co-training algorithm to classify human-rights abuses. In: 2017 International conference on the frontiers and advances in data science (FADS), 2017, pp 108–113

Gong C, Zhang H, Yang J, Tao D (2017) Learning with inadequate and incorrect supervision. In: 2017 IEEE international conference on data mining (ICDM), 2017, pp 889–894

Goodfellow I (2017) NIPS 2016 tutorial: generative adversarial networks. https://doi.org/10.48550/arXiv.1701.00160

Graef R (2021) Leveraging text classification by co-training with bidirectional language models—a novel hybrid approach and its application for a German bank. In: Innovation through information systems. WI 2021. Lecture notes in information systems and organisation, vol 47. Springer, pp 216–231

Guellil I, Adeel A, Azouaou F, Benali F, Hachani AE, Dashtipour K, Gogate M, Ieracitano C, Kashani R, Hussain A (2021) A semi-supervised approach for sentiment analysis of Arab(ic + izi) messages: application to the Algerian dialect. SN Comput Sci 2:118

Guo X, Wang W (2018) Towards making co-training suffer less from insufficient views. Front Comput Sci 13:99–105

Guo Q, Huang W (Wayne), Huang K, Liu X (2016) Information credibility: a probabilistic graphical model for identifying credible influenza posts on social media. In: Smart Health—international conference, ICSH 2015, revised selected papers, lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), 2016. Springer, pp 131–142. ISBN 9783319291741

Gupta R, Sahu S, Espy-Wilson C, Narayanan S (2018) Semi-supervised and transfer learning approaches for low resource sentiment classification. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2018, pp 5109–5113

Guru DS, Suhil M, Gowda HS, Raju LN (2016) Detection of a new class in a huge corpus of text documents through semi-supervised learning. In: 2016 International conference on advances in computing, communications and informatics (ICACCI), 2016, pp 494–499

Ha Q-T, Pham T-N, Nguyen V-Q, Nguyen M-C, Pham T-H, Nguyen T-T (2018a) A new text semi-supervised multi-label learning model based on using the label-feature relations. In: ICCCI, 2018

Han Y, Liu Y, Jin Z (2020) Sentiment analysis via semi-supervised learning: a model based on dynamic threshold and multi-classifiers. Neural Comput Appl 32(9):5117–5129

Hanafy M, Khalil MI, Abbas HM (2018) Combining classical and deep learning methods for Twitter sentiment analysis. In: ANNPR, 2018

Ha Q, Pham A, Nguyen VQ, Nguyen C, Vuong T-H, Tran M-T, Nguyen T-T (2018b) A new lifelong topic modeling method and its application to Vietnamese text multi-label classification. In: Intelligent information and database systems. ACIIDS 2018. Lecture notes in computer science, 2018, vol 10751, pp 200–210. ISBN 978-3-319-75416-1

Hartley HO, Rao JNK (1968) Classification and estimation in analysis of variance problems. Rev l'Inst Int Stat 36(2):141–147

Hasan A, Levene M, Weston D (2020) Learning structured medical information from social media. J Biomed Inform 110:103568. ISSN 1532-0464

Hassani H, Beneki C, Unger S, Mazinani MT, Yeganegi MR (2020) Text mining in big data analytics. Big Data Cogn Comput 4(1):1

He C, Peng L, Le Y, He J, Zhu X (2019) SECaps: a sequence enhanced capsule model for charge prediction. In: Artificial neural networks and machine learning—ICANN 2019: text and time series. Springer, Cham, pp 227–239. ISBN 978-3-030-30490-4

Hidetaka I, Wang Y (2019) A semi-supervised approach for identification of the sections in charge of RFQ documents. In: 2019 IEEE international conference on big data, 2019, pp 5532–5535

Hu W, Chen C, Chang Y, Zheng Z, Du Y (2021) Robust graph convolutional networks with directional graph adversarial training. Appl Intell 51:7812–7826

Huang J, Zhou Z, Shang J, Niu C (2020) Heterogeneous domain adaptation with label and structural consistency. Multimed Tools Appl 79:07

Huang J, Tao N, Chen H, Deng Q, Wang W, Wang J (2021) Semi-supervised text classification based on graph attention neural networks. In: 2021 4th International conference on artificial intelligence and big data (ICAIBD), 2021, pp 325–330

Huang L, Yu J, Hu Y, Chang H (2020a) A semi-supervised learning framework for TRIZ-based Chinese patent classification. In: Proceedings of the 2020 6th international conference on computing and artificial intelligence, ICCAI '20, 2020, pp 46–50. ISBN 9781450377089

Iglesias E, Vieira S, Diz LB (2016) An HMM-based multi-view co-training framework for single-view text corpora. In: Hybrid artificial intelligent systems. HAIS 2016. Lecture notes in computer science, 2016, vol 9648, pp 66–78. ISBN 978-3-319-32033-5

Jahanbakhsh Z, Feizi-Derakhshi MR, Sharifi A (2020) A semi-supervised model for Persian rumor verification based on content information. Multimed Tools Appl 80:1–29

Jedrzejowicz J, Zakrzewska M (2020) Text classification using LDA-W2V hybrid algorithm. In: Intelligent decision technologies 2019. Smart innovation, systems and technologies, vol 142, pp 227–237. ISBN 978-981-13-8310-6

Ji Y, Wang J, Niu Y, Ma H (2021) Reliable event detection via multiple edge computing on streaming traffic social data. IEEE Access. https://doi.org/10.1109/ACCESS.2021.3060624

Jiang M, Liang Y, Feng X, Fan X, Pei Z, Xue Y, Guan R (2018) Text classification based on deep belief network and softmax regression. Neural Comput Appl 29:01

Jing L (2018) Online fake comments detecting model based on feature analysis. In: 2018 International conference on smart grid and electrical automation (ICSGEA), 2018, pp 412–415

Ju W, Yang J, Qu M, Song W, Shen J, Zhang M (2022) KGNN: harnessing kernel-based networks for semi-supervised graph classification. In: Proceedings of the fifteenth ACM international conference on web search and data mining, WSDM '22, 2022, pp 421–429. ISBN 9781450391320

Kadhim AI (2019) Survey on supervised machine learning techniques for automatic text classification. Artif Intell Rev 52(1):273–292

Kang M, Biswas A, Kim D-C, Gao J (2019) Semi-supervised discriminative transfer learning in cross-language text classification. In: 2019 18th IEEE international conference on machine learning and applications (ICMLA), 2019, pp 1031–1038

Karisani P, Karisani N (2021) Semi-supervised text classification via self-pretraining. In: Conference: WSDM '21: the fourteenth ACM international conference on web search and data mining, 2021, pp 40–48. ISBN 9781450382977

Khan FH, Qamar U, Bashir S (2017) A semi-supervised approach to sentiment analysis using revised sentiment strength based on SentiWordNet. Knowl Inf Syst 51(3):851–872. ISSN 0219-1377

Khan A, Zubair M (2020) Classification of multi-lingual tweets, into multi-class model using Naïve Bayes and semi-supervised learning. Multimed Tools Appl 79:11

Kihlman R, Fasli M (2021) Classifying human rights violations using deep multi-label co-training. In: 2021 IEEE international conference on big data (Big Data), 2021, pp 4887–4895

Kontonatsios G, Brockmeier AJ, Przybyła P, McNaught J, Mu T, Goulermas JY, Ananiadou S (2017) A semi-supervised approach using label propagation to support citation screening. J Biomed Inform 72:67–76. ISSN 1532-0464

Kowsari K, Meimandi KJ, Heidarysafa M, Mendu S, Barnes L, Brown D (2019) Text classification algorithms: a survey. Information 10(4):150

Krishnamoorthy A, Patil AK, Vasudevan N, Pathari V (2018) News article classification with clustering using semi-supervised learning. In: 2018 International conference on advances in computing, communications and informatics (ICACCI), 2018, pp 86–91

Kumar T, Park J, Ali MS, Shahab Uddin AFM, Ko JH, Bae S-H (2021) Binary-classifiers-enabled filters for semi-supervised learning. IEEE Access 9:167663–167673

Lee VLS, Gan KH, Tan TP, Abdullah R (2019) Semi-supervised learning for sentiment classification using small number of labeled data. Procedia Comput Sci 161:577–584

Lee S, Kim W (2017) Sentiment labeling for extending initial labeled data to improve semi-supervised sentiment classification. Electron Commer Rec Appl 26(C):35–49. ISSN 1567-4223

Li AH, Sethy A (2020) Semi-supervised learning for text classification by layer partitioning. In: ICASSP 2020—2020 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2020, pp 6164–6168

Li Y, Su L, Chen J, Yuan L (2017) Semi-supervised learning for question classification in CQA. Natural Comput 16:12

Li Z, Yang F, Luo Y (2019) Context embedding based on bi-LSTM in semi-supervised biomedical word sense disambiguation. IEEE Access 7:72928–72935

Li M, Dai Q (2018) A novel knowledge-leverage-based transfer learning algorithm. Appl Intell 48(8):2355–2372. ISSN 0924-669X

Lieder I, Segal M, Avidan E, Cohen A, Hope T (2019) Learning a faceted customer segmentation for discovering new business opportunities at Intel. In: 2019 IEEE International conference on big data (Big Data), 2019, pp 6136–6138

Li M, Lang C, Yu M, Lu Y, Liu C, Jiang J, Huang W (2020) SCX-SD: semi-supervised method for contextual sarcasm detection. In: Knowledge science, engineering and management, 2020. Springer, Cham, pp 288–299. ISBN 978-3-030-55393-7

Li W, Li Y, Chen J, Hou C (2017) Product functional information based automatic patent classification: method and experimental studies. Inf Syst 67:71–82. ISSN 0306-4379

Lin J, Mao W, Zeng D (2017) Topic and user based refinement for competitive perspective identification. In: IEEE international conference on intelligence and security informatics (ISI), 2017, pp 131–133

Linmei H, Yang T, Shi C, Ji H, Li X (2019) Heterogeneous graph attention networks for semi-supervised short text classification. In: Proceedings of the 2019 conference on empirical methods in natural language processing (EMNLP) and the 9th international joint conference on natural language processing, 2019, pp 4821–4830

Li L, Qiu X (2020) TAVAT: token-aware virtual adversarial training for language understanding. https://doi.org/10.48550/arXiv.2004.14543

Liu C-L, Hsaio W-H, Lee C-H, Chang T-H, Kuo T-H (2016) Semi-supervised text classification with universum learning. IEEE Trans Cybern 46(2):462–473

Liu J, Timsina P, El-Gayar O (2018) A comparative analysis of semi-supervised learning: the case of article selection for medical systematic reviews. Inf Syst Front 20:04

Liu J, Deng J, Xu G, He Z (2018a) In: Hierarchical attention based semi-supervised network representation learning: 7th CCF international conference, NLPCC 2018, Hohhot, China, 26–30 August 2018, proceedings, Part I, pp 237–249. ISBN 978-3-319-99494-9

Liu L, Li Y, Xiong Y, Cavallucci D (2020) A new function-based patent knowledge retrieval tool for conceptual design of innovative products. Comput Ind 115:103154. ISSN 0166-3615

Liu X, Long F, Huang K, Ling Q (2021) Enhanced unsupervised data augmentation for emergency events detection and classification. In: 33rd Chinese control and decision conference, 2021, pp 2367–2371

Li X, Yan L, Qin N, Ran H (2017a) A novel semi-supervised short text classification algorithm based on fusion similarity. In: Intelligent computing methodologies, 2017. Springer, Cham, pp 309–319. ISBN 978-3-319-63315-2

Li Y, Ye J (2018) Learning adversarial networks for semi-supervised text classification via policy gradient. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery; data mining, KDD 18, 2018, pp 1715–1723. ISBN 9781450355520

Li P, Zhao F, Li Y, Zhu Z (2018) Law text classification using semi-supervised convolutional neural networks. In: 2018 Chinese control and decision conference (CCDC), 2018, pp 309–313

Ma B, Sun H, Wang J, Qi Q, Liao J (2020) Semi-supervised sentence classification based on user polarity in the social scenarios. In: ICASSP 2020—2020 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2020, pp 8209–8213

Mathapati S, Nafeesa A, Tanuja R, Manjula SH, Venugopal KR (2019) Semi-supervised domain adaptation and collaborative deep learning for dual sentiment analysis. SN Appl Sci 1:907

McNulty J, Alvarez S, Langmayr M (2021) Detecting research from an uncurated HTML archive using semi-supervised machine learning. In: 2021 Systems and information engineering design symposium (SIEDS), 2021, pp 1–6

Minaee S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J (2021) Deep learning-based text classification: a comprehensive review. ACM Comput Surv 54(3):1–40

Miyato T, Dai AM, Goodfellow I (2017) Adversarial training methods for semi-supervised text classification. In: Conference paper at ICLR 2017, 2017

Moon S, Carbonell J (2016) Proactive transfer learning for heterogeneous feature and label spaces. In: Machine learning and knowledge discovery in databases. Springer, Cham, pp 706–721. ISBN 978-3-319-46227-1

Najari S, Salehi M, Farahbakhsh R (2022) GANBOT: a GAN-based framework for social bot detection. Soc Netw Anal Min 12:4

Namrutha Sridhar BV, Mrinalini K, Vijayalakshmi P (2020) Data annotation and multi-emotion classification for social media text. In: 2020 International conference on communication and signal processing (ICCSP), 2020, pp 1011–1015

Ng LHX, Carley KM (2021) "The coronavirus is a bioweapon'': classifying coronavirus stories on fact-checking sites. Comput Math Organ Theory 27(2):179–194

Nguyen M (2016) Leveraging emotional consistency for semi-supervised sentiment classification. In: Advances in knowledge discovery and data mining. PAKDD 2016. Lecture notes in computer science, 2016, vol 9651, pp 369–381. ISBN 978-3-319-31752-6

Nguyen-Nhat DK, Duong H-T (2019) One-document training for Vietnamese sentiment analysis. In: Computational data and social networks. CSoNet 2019. Lecture notes in computer science, 2019, vol 11917, pp 189–200. ISBN 978-3-030-34979-0

Nigam K, McCallum A, Thrun S, Mitchell T (2000) Text classification from labeled and unlabeled documents using EM. Mach Learn 39:103–134. https://doi.org/10.1023/A:1007692713085

Omar A, Mahmoud TM, Abd-El-Hafeez T, Mahfouz A (2021) Multi-label Arabic text classification in online social networks. Inf Syst 100:101785. ISSN 0306-4379

Pan Y, Chen Z, Suzuki Y, Fukumoto F, Nishizaki H (2020) Sentiment analysis using semi-supervised learning with few labeled data. In: 2020 International conference on cyberworlds (CW), 2020, pp 231–234

Pang B, Lee L (2005) Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the 43rd annual meeting of the Association for Computational Linguistics (ACL'05), 2005. Association for Computational Linguistics, Ann Arbor, pp 115–124

Park S, Lee J, Kim K (2019) Semi-supervised distributed representations of documents for sentiment analysis. Neural Netw 119:139–150. ISSN 0893-6080

Pavlinek M, Podgorelec V (2017) Text classification method based on self-training and LDA topic models. Expert Syst Appl 80:83–93. ISSN 0957-4174

Pohl M, Hashaam A, Bosse S, Staegemann DG, Volk M, Kramer F, Turowski K (2020) Application of NLP to determine the state of issues in bug tracking systems. In: 2020 International conference on data mining workshops (ICDMW), 2020, pp 53–61

Qiu Y, Gong X, Ma Z, Chen X (2020) MixLab: an informative semi-supervised method for multi-label classification. In: Natural language processing and Chinese computing, 2020. Springer, Cham, pp 506–518. ISBN 978-3-030-60450-9

Rossi R, Lopes A, Rezende S (2017) Using bipartite heterogeneous networks to speed up inductive semi-supervised learning and improve automatic text categorization. Knowl-Based Syst 132:06

Sajeeda A, Mainul Hossain BM (2022) Exploring generative adversarial networks and adversarial training. Int J Cogn Comput Eng 3:78–89. ISSN 2666-3074. https://doi.org/10.1016/j.ijcce.2022.03.002

Sakai T, Niu G, Sugiyama M (2017) Semi-supervised AUC optimization based on positive-unlabeled learning. https://doi.org/10.48550/arXiv.1705.01708

Severin K, Gokhale S, Dagnino A (2019) Keyword-based semi-supervised text classification. In: 2019 IEEE 43rd annual computer software and applications conference (COMPSAC), 2019, vol 1, pp 417–422

Shahri MP, Roe MM, Reynolds G, Kahanda I (2019) PPPred: classifying protein-phenotype co-mentions extracted from biomedical literature. bioRxiv

Shayegh P, Li Y, Zhang J, Zhang Q (2019) Semi-supervised text classification with deep convolutional neural network using feature fusion approach. In: 2019 IEEE/WIC/ACM international conference on web intelligence (WI), 2019, pp 363–366

Shehnepoor S, Togneri R, Liu W, Bennamoun M (2022) ScoreGAN: a fraud review detector based on regulated GAN with data augmentation. IEEE Trans Inf Forensics Secur 17:280–291
Shulman H, Simo H (2021) Poster: WallGuard—a deep learning approach for avoiding regrettable posts in social media. In: 2021 IEEE 41st international conference on distributed computing systems (ICDCS), 2021, pp 1142–1143
Soleimani H, Miller DJ (2016a) Exploiting the value of class labels in topic models for semi-supervised document classification. In: International joint conference on neural networks, 2016, pp 4025–4031
Soleimani H, Miller DJ (2016b) Semi-supervised multi-label topic models for document classification and sentence labeling. In: Proceedings of the 25th ACM international on conference on information and knowledge management, CIKM '16, 2016, pp 105–114. ISBN 9781450340731
Song H-J, Park S-B (2018) Identifying intention posts in discussion forums using multi-instance learning and multiple sources transfer learning. Soft Comput 22:12
Song J, Qin S, Zhang P (2016) Chinese text categorization based on deep belief networks. In: 2016 IEEE/ACIS 15th international conference on computer and information science, 2016, pp 1–5
Stanojevic M, Alshehri J, Obradovic Z (2019) Surveying public opinion using label prediction on social media data. In: Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining, ASONAM '19, 2019, pp 188–195. ISBN 9781450368681
Stanton G, Irissappane AA (2019) GANs for semi-supervised opinion spam detection. https://doi.org/10.48550/arXiv.1903.08289
Statista (2022) Internet user growth worldwide from 2018 to 2023. https://www.statista.com/statistics/1190263/internet-users-worldwide/
Steyn C, de Waal A (2016) Semi-supervised machine learning for textual anomaly detection. In: 2016 Pattern Recognition Association of South Africa and robotics and mechatronics international conference (PRASA-RobMech), 2016, pp 1–5
Sukhija S, Krishnan NC (2019) Web-induced heterogeneous transfer learning with sample selection. In: Machine learning and knowledge discovery in databases, 2019. Springer, Cham, pp 777–793. ISBN 978-3-030-10928-8
Sun L, Ge H, Kang W (2018) Non-negative matrix factorization based modeling and training algorithm for multi-label learning. Front Comput Sci 13:11
Sun K, Lin Z, Guo H, Zhu Z (2019b) Virtual adversarial training on graph convolutional networks in node classification. In: Pattern recognition and computer vision, 2019. Springer, Cham, pp 431–443. ISBN 978-3-030-31654-9
Sun C, Qiu X, Xu Y, Huang X (2019a) How to fine-tune BERT for text classification? In Chinese computational linguistics. Springer, Cham, pp 194–206. ISBN 978-3-030-32381-3
Sun Z, Zhang X, Ye Y, Chu X, Liu Z (2020) A probabilistic approach towards an unbiased semi-supervised cluster tree. Knowl-Based Syst 192:105306. ISSN 0950-7051
Tanha J (2018) MSSBoost: a new multiclass boosting to semi-supervised learning. Neurocomputing 314:251–266. ISSN 0925-2312
Tanha J (2019) A multiclass boosting algorithm to labeled and unlabeled data. Int J Mach Learn Cybern 10:12
Thangaraj M, Sivakami M (2018) Text classification techniques: a literature review. Interdiscip J Inf Knowl Manag 13:117
Thomas A, Resmipriya MG (2016) An efficient text classification scheme using clustering. Procedia Technol 24:1220–1225
Timsina P, Liu J, El-Gayar O, Shang Y (2016) Using semi-supervised learning for the creation of medical systematic review: an exploratory analysis. In: 2016 49th Hawaii international conference on system sciences (HICSS), 2016, pp 1195–1203
Tollefson J (2018) China declared world's largest producer of scientific articles. Nature 553:390–390
van Engelen JE, Hoos HH (2019) A survey on semi-supervised learning. Mach Learn 109:373–440
Van Engelen JE, Hoos HH (2020) A survey on semi-supervised learning. Mach Learn 109(2):373–440
Varghese A, Cawley M, Hong T (2018) Supervised clustering for automated document classification and prioritization: a case study using toxicological abstracts. Environ Syst Decis 38:09
Vilhagra LA, Fernandes ER, Nogueira BM (2020) TextCSN: a semi-supervised approach for text clustering using pairwise constraints and convolutional Siamese network. In: SAC '20: proceedings of the 35th annual ACM symposium on applied computing, 2020, pp 1135–1142. ISBN 9781450368667
Villatoro-Tello E, Anguiano E, Montes M, Villaseñor-Pineda L, Ramirez-de-la Rosa G (2016) Enhancing semi-supervised text classification using document summaries. In: Advances in artificial intelligence—IBERAMIA 2016. Lecture notes in computer science, 2016, vol 10022, pp 115–126. ISBN 978-3-319-47954-5

Wang W, Tan G, Wang H (2017) Cross-domain comparison of algorithm performance in extracting aspect-based opinions from Chinese online reviews. Int J Mach Learn Cybern 8:06

Wang Y, Gu Q, Brown D (2019) Differentially private hypothesis transfer learning. In: Machine learning and knowledge discovery in databases. Springer, pp 811–826. ISBN 978-3-030-10928-8

Wang X, Ren J (2019) Semi-supervised learning for classification on Chinese drug treatment questions. In: 2019 IEEE international conference on bioinformatics and biomedicine, 2019, pp 991–994

Wang Z, Tu E, Lee Z (2021) Deep semi-supervised learning via dynamic anchor graph embedding learning. In: 2021 International joint conference on neural networks (IJCNN), 2021, pp 1–8

Widmann N, Verberne S (2017) Graph-based semi-supervised learning for text classification. In: Proceedings of the ACM SIGIR international conference on theory of information retrieval, 2017, pp 59–66. ISBN 9781450344906

Wu F, Jing X-Y, Zhou J, Ji Y, Lan C, Huang Q, Wang R (2019) Semi-supervised multi-view individual and sharable feature learning for webpage classification. In: WWW '19, 2019, pp 3349–3355. ISBN 9781450366748

Wulan SR, Supangkat SH (2017) Semi-supervised learning self-training for Indonesian motivational messages classification. In: 2017 International conference on ICT for smart society, 2017, pp 1–7

Xiang R, Yin S (2021) Semi-supervised text classification with temporal ensembling. In: 2021 International conference on computer communication and artificial intelligence (CCAI), 2021, pp 204–208

Xiao H, Liu X, Song Y (2019) Efficient path prediction for semi-supervised and weakly supervised hierarchical text classification. In: The World Wide Web conference on—WWW '19, 2019

Xie Q, Huang J, Peng M, Zhang Y, Peng K, Wang H (2019) Discriminative regularized deep generative models for semi-supervised learning. In: 2019 IEEE international conference on data mining (ICDM), 2019, pp 658–667

Xu X, Li W, Xu D, Tsang IW (2016) Co-labeling for multi-view weakly labeled learning. IEEE Trans Pattern Anal Mach Intell 38(6):1113–1125

Xu Z, Li J, Liu B, Bi J, Li R, Mao R (2017) Semi-supervised learning in large scale text categorization. J Shanghai Jiaotong Univ (Sci) 22:291–302

Xu B, Huang J, Hou L, Shen H, Gao J, Cheng X (2020) Label-consistency based graph neural networks for semi-supervised node classification. In: SIGIR '20: the 43rd international ACM SIGIR conference on research and development in information retrieval, 2020, pp 1897–1900

Xu Y, Li B (2017) Sentiment classification incorporating user profile. In: 2017 4th International conference on information science and control engineering (ICISCE), 2017, pp 663–667

Yadav M, Bhojane V (2019) Semi-supervised mix-Hindi sentiment analysis using neural network. In: 9th International conference on cloud computing, data science engineering, 2019, pp 309–314

Yadav S, Kumar G, Kumar S (2019) A graph construction study for graph-based semi-supervised learning: case study on unstructured text data. In: International conference on Big Data, 2019, pp 6254–6256

Yang F, Zhang H, Tao S (2021) Simplified multilayer graph convolutional networks with dropout. Appl Intell 52:4776–4791

Yang T, Linmei H, Shi C, Ji H, Li X, Nie L (2021a) HGAT: heterogeneous graph attention networks for semi-supervised short text classification. 39(3). ISSN 1046-8188

Yin Z, Xiang J, Yin C, Wang J (2018) Text classification algorithm based on SLAS-C. In: Advances in computer science and ubiquitous computing. CUTE CSA 2017. Lecture notes in electrical engineering, 2018, vol 474, pp 382–387. ISBN 978-981-10-7604-6

Yu X, Ren C, Zhou Y, Wang Y (2016) A transductive support vector machine algorithm based on ant colony optimization. In: Social computing. ICYCSEE 2016. Communications in computer and information science, vol 623, pp 127–135. ISBN 978-981-10-2052-0

Yu J, Wu J, Wei B, Liu Y (2019) CVAE-attention: CVAE based semi-supervised sentiment classification using attention. In: Proceedings of the 2019 international conference on pattern recognition and artificial intelligence, PRAI '19, 2019, pp 68–75. ISBN 9781450372312

Zaghdoudi S, Glomann L (2021) Artificial intelligence enabled user experience research. In: Advances in artificial intelligence, software and systems engineering, pp 187–193. ISBN 978-3-030-51327-6

Zhang Y, Ma J, Wang Z (2019) Semi supervised classification of scientific and technical literature based on semi supervised hierarchical description of improved latent Dirichlet allocation (LDA). Clust Comput 22:05

Zhang W, Chen Q, Chen Y (2020) Deep learning based robust text classification method via virtual adversarial training. IEEE Access 8:61174–61182

Zhang Z, Luo J, Huang G (2019b) A semi-supervised short text classification method based on weighted word vector representation. In: 2019 IEEE 9th international conference on electronics information and emergency communication (ICEIEC), 2019, pp 324–329

Zhang X, Zhang C, Luna DX, Shang J, Han J (2021b) Minimally-supervised structure-rich text categorization via learning on text-rich networks. In: Proceedings of the web conference 2021, WWW '21, 2021, pp 3258–3268. ISBN 9781450383127

Zhang G, Zheng H, Liu XY (2021a) Co-STM text categorization method based on supervised topic model. In: 2021 4th International conference on advanced electronic materials, computers and software engineering (AEMCSE), 2021, pp 462–467

Zhao H, Xie J, Wang H (2022) Graph convolutional network based on multi-head pooling for short text classification. IEEE Access 10:11947–11956

Zhou Z-H (2021) Semi-supervised learning. In: Machine learning. Springer, Berlin, pp 315–341

Zhu W, Liu Y, Hu G, Ni J, Lu Z (2018) A sample extension method based on Wikipedia and its application in text classification. Wirel Pers Commun 102:10

Zhu D-H, Dai X-Y, Chen J-J (2021) Pre-train and learn: preserving global information for graph neural networks. J Comput Sci Technol 36(6):1420–1430