DOI: 10.6977/IJoSI.202403_8(1).0002
O. Stitini, S. Kaloun, etc./Int. J. Systematic Innovation, 8(1), 11-26 (2024)

An improved self-training model to detect fake news categories using multi-class classification of unlabeled data: fake news classification with unlabeled data

Oumaima Stitini[1]*, Soulaimane Kaloun[2], Omar Bencharef[3], Sara Qassimi[4]
[1] Computer and system engineering laboratory, ENS, Cadi Ayyad University, Marrakesh, 40000, Morocco.
[2,3,4] Computer and system engineering laboratory, FSTG, Cadi Ayyad University, Marrakesh, 40000, Morocco.
* Corresponding author E-mail:o.stitini@uca.ac.ma, oumaima.stitini@ced.uca.ma

(Received 05 May 2023; Final version received 18 December 2023; Accepted 23 December 2023.)

Abstract

In recent times, significant attention has been devoted to classifying news content in academic and industrial settings. Some studies have focused on distinguishing between fake and real news using labeled data and have achieved some success in detection. Digital misinformation or fake news content spreads through online social communities via shares, re-shares, and re-posts. Social media has faced several challenges in combating the distribution of fake news information. Social media platforms and blogs have become widely used daily sources of information due to their low cost and ease of access. However, this widespread use of social media for news consumption has led to the dissemination of fake news, creating a severe problem that adversely affects individuals and society. Consequently, identifying and addressing misinformation has become an essential and critical task. Detecting fake news is an emerging research area that has garnered considerable interest, but it also presents specific challenges, mainly due to the limitations of available resources. In this paper, we focus on identifying and classifying different forms of fake news using unlabeled data, specifically exploring how to use unlabeled data for multi-class classification. The proposed approach categorizes fake news into four forms: satire or fake satirical information, manufacturing, manipulation, and propaganda. Our method employs a relevant approach based on multi-class classification using unlabeled data. The experimental evaluation demonstrates the efficiency of our suggested system.

Keywords: Multi-class classification, Unlabeled data, Semi-supervised learning, Self-training, Recommender system, Fake news, Imbalanced Learning

1. Introduction

It is certainly critical to identify and mitigate fake news, representing a challenging and socially relevant of fake news has opened up new academic directions, conducting challenging studies to counter the problem. Many research studies have focused on identifying and containing fake news through mitigation techniques (Qian et al., 2018). Digital misinformation or fake news content is spread through social communities via shares, re-shares, and re-posts. The spread of this misinformation through social networks follows a similar pattern to the transmission of infectious diseases. Therefore, insights about the spread of fake news can be gained from analyzing the dynamics of transmission. For example, the recent coronavirus pandemic, causing COVID-19, can evolve and compete in a host population shaped by social contacts, much like rumors and fake news. The propagation of information on social media is inundated with fake news, taking different forms. Some express humor, while others are serious and create doubt in the public (Collins et al., 2020).

The identification of misleading information involves determining the truthfulness of news by examining its content and related information, such as dissemination patterns. This issue has garnered significant interest from various perspectives, with supervised learning being the dominant approach for fake news identification, which has achieved success. Many research efforts aim to detect fake news using labeled data. Different studies focus on classifying fake news on social media, targeting various types of fake news. Some studies concentrate on distinguishing between fake and real news (Vijayaraghavan, 2020), while others focus on a specific type of fake news (Li et al., 2019; Alzanin & Azmi, 2018). Certain works are dedicated to classifying two or three types of rumors (Wang, 2017), such as early detection of rumors (Wu et al., 2018) or curbing their spread (Imran et al., 2015). Thus, our objective is to detect different forms of fake news in the absence of labeled data. In this paper, we investigate the identification of fake news with varying degrees of fakeness by leveraging multiple sources. We address the problem of multi-class classification for detecting fake news forms using unlabeled data. In particular, we aim to answer four primary research questions mentioned in Fig. 1. The key contributions of this paper are as follows: Section 2 contains our literature review and theoretical background. We describe our state-of-the-art in Section 3. Then, in Section 4, we will elaborate on the proposed approach. In Section 5, we mention our experiments and results. At the end of the work, we discuss and conclude all the work in Section 6.

Table 1: Research Question

No	Research Question
RQ1	How to differentiate between fake news forms?
RQ2	How to effectively detect the right form of fake news using multi-class classification?
RQ3	How to efficiently perform multi-class classification using unlabeled data?
RQ4	How to improve self-training algorithms for multi-class classification?

2. Literature review and theoretical background

In this section, we review various research works (Oumaima et al., 2020) about the detection of fake news. Most research has tackled this problem using supervised learning algorithms. In natural language processing, detecting fake news necessitates a substantial amount of labeled data to build effective detection models through supervised learning. However, recording information from social media is prohibitively expensive and demands significant human effort due to the sheer volume of social media data. As data grows exponentially, relying solely on labeled data becomes impractical for enhancing fake news detection. Therefore, exploring solutions that leverage unlabeled data to improve detection and address this limitation holds promise (Tanha, 2019).

2.1 Different forms of fake news

Research works on fake news are at an early stage and require deep analysis to precisely choose the relevant features. In general, we could categorize fake news into two levels as we did in our previous research work (Stitini et al., 2022):

1. High level:
   - ❖ **Manufacturing**: Involves the creation of false information in newspapers or other media sources to gain credibility and deceive the audience.
   - ❖ **Manipulation**: Involves the deceptive alteration of images or videos, removing them from their original context to spread false news.
   - ❖ **Propaganda**: Aims to influence public opinion and modify people's perception of events to serve a particular agenda.
2. Low level:
   - ❖ **Satire** or false satirical information: Designed primarily to provide humor to readers but may be mistaken as genuine news.
   - ❖ **Parody**: A comedic form that uses the structure, characters, style, and functioning of a work or institution to mock it.

Among the types of information that are likely to be fake news, we quote:

- ➢ **Pure information**: A presentation of facts without any analysis by the journalist, potentially lacking context.
- ➢ **Described information**: The facts are described in relation to a specific social or psychological behavior, which may lead to misleading interpretations.
- ➢ **Analyzed information**: The facts are analyzed, connecting them to past events or projecting potential future outcomes, possibly leading to biased interpretations.
- ➢ **Commented information**: Involves value judgments on the presented facts, which could skew the perception of the news.

2.2 Types of classification

Classification is a fundamental task in machine learning and data analysis. It involves categorizing data points into predefined classes or groups based on their features or attributes. There are several types of classification problems, including:

- ❖ Binary Classification: This type of classification involves dividing data into two distinct classes or categories. For example, determining whether an email is spam or not spam, predicting whether a patient has a particular disease or not, etc.
- ❖ Multi-class Classification: In multi-class classification, data points are classified into three or more categories. Each instance belongs to one and only one class. For instance, classifying animals into categories like mammals, birds, reptiles, etc (Stitini et al., 2022a, Kaliyar et al., 2019).
- ❖ Multi-label Classification: In multi-label classification, data points can be associated with multiple classes or labels simultaneously. For instance, tagging an image with multiple objects or identifying topics in a document with various labels. In recent years, multi-label classification has become very relevant because of its vast range of implementation areas; each input sample is identified with target objects in a multi-label classification. The number of ticket labels associated with each entry is unknown; it varies dynamically (Rasool et al., 2019b). Several methods have been developed for multi-label classification: Algorithm Adaptation Methods, Problem Transformation Methods, and Ensemble Methods.
- ❖ Imbalanced Classification: Imbalanced classification occurs when the distribution of classes in the dataset is highly skewed, meaning that one class has significantly more instances than others. Handling imbalanced data is a challenging problem in classification. The unequal class distribution can be named as an imbalanced classification and defined by the ratio of the majority of individuals who belong to the minority class to that of the majority class. One of the critical issues of imbalanced classification is simultaneous class occurrences in datasets (Jedrzejowicz et al., 2018). There are two strategies to handle class in general, and there are two methods for dealing with class imbalance classification: 1) data level approach and 2) algorithm level approach. Methods on the data level approach change the imbalanced class ratio to obtain a balanced division between classes. Simultaneously, standard classification algorithms are set on the algorithm level approach to increase the learning task speed.

Each instance of the learning set belongs to a series of label sets previously defined in several classifications. There are three types of approaches for dealing with multiple-class classification problems.

1. Extension of the binary case: Different algorithms based on support vector machines, naive Bayes, neural network decisions, Neighbors, and extreme learning machines are designed to solve multi-class classification problems.
2. Conversion of the multi-class classification problem into several binary classification problems: It reduces the problem of multi-class classification to multiple binary classification issues. It can be classified in One Vs Rest and One vs One.
    a. One-vs-Rest (OvR) or One-vs-All (OvA): In OvR, each class is treated as the positive class, and the remaining classes are treated as the negative class for separate binary classifiers. Each classifier predicts whether an instance belongs to the positive class or not.
    b. One-vs-One (OvO): In OvO, a separate binary classifier is trained for each pair of classes. The class with the most votes from all classifiers is selected as the final prediction.
3. Hierarchical classification methods: Hierarchical classification addresses the multi-class classification problem by dividing the output space in a tree. Each parent node is divided into several child nodes, and the process continues until each child node is only one class. Several approaches focused on hierarchical classification have been suggested, like Binary Hierarchical Classifiers, and Divide-By-2 (Silva-Palacios et al., 2017).

DOI: 10.6977/IJoSI.202403_8(1).0002
O. Stitini, S. Kaloun, etc./Int. J. Systematic Innovation, 8(1), 11-26 (2024)

2.3 Multi-class classification using unlabeled Data

Standard classifying algorithms use supervised learning, where the classifier is trained solely on labeled information. However, many real-world classification problems present complexities, costs, or time constraints, as they require observational studies. In contrast, obtaining unlabeled data is inexpensive and requires less effort from experienced individuals. Semi-supervised learning algorithms offer a suitable and scalable machine learning approach for utilizing labeled and unlabeled data to construct effective classifiers (Forestier & Wemmert, 2016, Larriva-Novo et al., 2020).

2.4 Vectorization of text data

Transforming text data into interactive vectors enables interactions with machines for solving mathematical problems and performing natural language processing tasks. Researchers in this field have proposed various vectorization models, ranging from simple to elaborate, to address NLP challenges. Here is a brief introduction to standard text vectorization methods and new word embedding models:

- **TF-IDF**: TF-IDF is the most common NLP approach for mapping text documents to matrix vectors. It represents the importance of a term in a collection of documents for a specific document. Convincing search engines can be built using future TF-IDF scores to capture prominent terms in the text, thus enhancing document relevance for specific search queries. However, the inverse document frequency (IDF) term's selectiveness limits the TF-IDF score's adaptability in handling dynamic uncertainties in text.
- **Word2Vec**: Word2Vec generates distributed semantic representations for words in a document. The model aims to develop each word's sense, resulting in similar digital representations for related words. Word2Vec is
a predictive model that learns vectors to predict target terms based on their contextual word.
- **SentenceToVec**: SentenceToVec is an extension of Word2Vec, where vector representations of words in a sentence are averaged to learn character representations at the sentence level or for a full text. Skip-Thought Vectors, published in 2015, has significantly advanced sentence-level embeddings.
- **Doc2Vec**: Doc2Vec is an extension of Word2Vec, or SentenceToVec, as sentences are part of documents. The process for acquiring Doc2Vec embeddings is similar to that of SentenceToVec.

2.5 Statement based similarity methods

Term-based similarity measures can be divided into the following:

1. Cosine Similarity: Cosine similarity utilizes the angle between two vectors in an inner product space to determine their similarity.
2. Euclidean distance or L2 distance: This measure is calculated as the square root of the sum of the squared differences between the corresponding elements of the two vectors.
3. Jaccard similarity: Jaccard similarity is computed as the ratio of shared terms to the total number of unique terms in both strings.

3. State of the art

This article reviews various studies on different multi-class classification approaches using unlabeled data. Each approach examines challenges and analyzes vital aspects. The section is divided into two subsections. The first outlines the steps for sorting articles, including defining keywords, setting inclusion and exclusion criteria, and specifying the databases searched. The second subsection compares and discusses various related works.

3.1. Procedure for systematic review

This section presents the papers selected based on the applied area and the methods used, summarizing the various steps taken to carry out this study.

- Step 1: Definition of research questions and keywords: The research questions (Table 1) and the keywords (Table 2) were defined in this step.
- Step 2: Choice of search sources: In this step, articles and chapters were selected from the Scopus and Web of Science databases due to their credibility, relevance to the computing rea, and publication in high-ranked journals, conferences, and books by reputable

publishers such as IEEE, Elsevier, ACM, and Springer.

- Step 3: Elaboration of inclusion and exclusion criteria: Criteria were developed to identify papers that would undergo a complete reading. These criteria (Table 3) are discussed as follows. The papers were either selected for full reading or excluded based on the inclusion and exclusion criteria.
- Step 4: Thorough reading of selected papers: The papers selected in Step 3 were thoroughly read and evaluated for their relevance to the research scope. Finally, the papers were ranked highly relevant, partially relevant, or irrelevant to the established research questions.

Table 2: Research Strings

No	Search Via Keywords
S1	"Multi Classification using unlabeled data" OR "Multi-class classification using unlabeled data"
S2	"Multi-class classification using unlabeled data" AND "self-training algorithm"
S3	"Fake news detection" AND "Multi-classification" OR "Multi-class classification with unlabeled data to detect fake news forms"

Table 3: Inclusion and exclusion criteria

Including Criteria	Excluding criteria
Paper was published in a journal as a scientific article.	Paper is not written in English.
Paper published from 2016 to 2020.	Paper published before 2016.
Indexed paper.	Not Indexed paper.

3.2 Research contribution

A. Motivation

In the era of big data, obtaining labeled data for every task, such as classifying news as fake or real, can be challenging and time-consuming. Incorporating unlabeled data and using self-training is a powerful approach to mitigate this issue and improve classification performance.

Self-training is a semi-supervised learning technique where a model iteratively labels unlabeled data and uses the newly labeled data to retrain itself, gradually improving its performance. The process typically involves the following steps:

★ **Initial Model training:** Start by training a model on the limited labeled data you have. This will be the initial version of your classifier.

★ **Pseudo-labeling**: Use the initial model to make predictions on the unlabeled data. Assign pseudo-labels (labels generated by the model) to the unlabeled samples based on their predicted classes.

★**Combine labeled and pseudo-labeled data:** Merge the labeled data and the newly pseudo-labeled data to form a larger training set.

★ **Retraining:** Retrain the model on this combined dataset. The model now has more data, including both labeled and pseudo-labeled examples.

★ **Repeat:** Iterate the pseudo-labeling and retraining process for a certain number of iterations or until convergence.

By incorporating unlabeled data through self-training, the model can learn from a more diverse and comprehensive dataset, which often leads to improved performance. However, it's essential to be cautious about potential noise in the pseudo-labeling process, as incorrect pseudo-labels can propagate and harm the model's performance. Techniques such as entropy-based filtering and using confidence thresholds can help reduce the impact of noisy pseudo-labels. Additionally, active learning can be employed in combination with self-training to intelligently select the most informative unlabeled samples for pseudo-labeling, further improving the efficiency of the process.

In conclusion, using self-training and incorporating unlabeled data can be a powerful solution to enhance the classification of news as fake or real when labeled data is limited or hard to obtain. However, it requires careful implementation and consideration of potential challenges like noisy pseudo-labeling

B. Research comparison

DOI: 10.6977/IJoSI.202403_8(1).0002
O. Stitini, S. Kaloun, etc./Int. J. Systematic Innovation, 8(1), 11-26 (2024)

(Deepti Nikumbh, et al., 2023) discusses the prevalence of fake news on social media, its impact, and the challenges in detecting it. It highlights state-of-the-art methods relying on news content, user profiles, and social context features. The importance of feature engineering and extraction for fake news detection is emphasized, along with the need for future research in this area. Our proposed approach for fake news detection seems comprehensive as it categorizes fake news into four distinct forms: satire or fake satirical information, manufacturing, manipulation, and propaganda. This categorization acknowledges the different types of misinformation and disinformation that can be present in news content.

Incorporating unlabeled data is an intelligent decision as it allows the model to learn from a more extensive and diverse dataset, which is beneficial in scenarios where labeled data might be limited and difficult to obtain. By employing unlabeled data, the model can potentially uncover patterns and structures within the data that were not evident before, leading to improved performance.

3.3 Related work

Most previous efforts have demonstrated that unlabeled data can significantly boost classification accuracy when too few labeled samples are available. These methods can be divided into three major groups, as shown in Fig. 3:

1. Semi-supervised approach.
2. Clustering approach.
3. Deep learning approach.

This study has three multi-class classification approaches (Table 9). The table compares related works already conducted on multi-class classification using unlabeled data.

3.3.1 Semi-supervised approach

Semi-supervised learning is often considered the safest and most effective approach when dealing with the absence of labeled data and an abundance of unlabeled data in the training process. The purpose of proposing a semi-supervised learning method is to enhance learning outcomes and address various problems based on different data types. Several algorithms have been developed recently, including self-labeled, semi-supervised boosting, margin-based, graph-based, and generative methods.

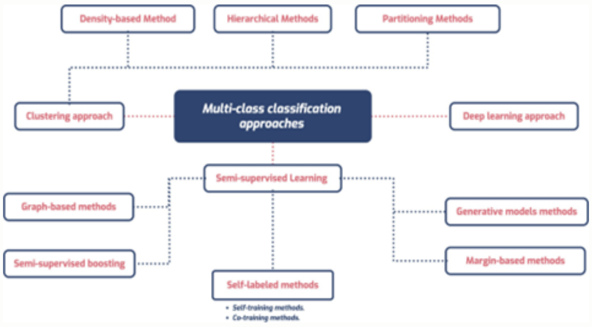

Fig. 1: Multi-class classification approaches.

In the context of multi-class classification, various approaches address different aspects. While these approaches are distinct, they are not mutually exclusive and can be combined to create more advanced multi-class classification systems. For instance, deep learning models can be trained on semi-supervised data to leverage unlabeled information, or clustering can be utilized as a preprocessing step to group similar instances before applying a classification algorithm. The choice of which approach or combination of approaches depends on the specific problem, the available data, and the desired performance. Semi-supervised learning is beneficial when labeled data is scarce or expensive. By leveraging the large amounts of unlabeled data, these techniques can improve model performance beyond what could be achieved with just labeled data. As with any machine learning approach, selecting the appropriate semi-supervised learning technique depends on the specific problem and the characteristics of the available data.

Self-labeled methods:

Self-labeled methods can be divided into two sub-categories: Self-training and Co-training.

- **Self-training methods:** Self-training is an iterative technique used in semi-supervised learning, considered among the primary models of repetitive strategies. Initially, a classifier is trained using labeled data. The classifier then assigns labels to each unlabeled data point, and the most trustworthy unlabeled points, along with their anticipated labels, are added to the training set. Existing work (D. Wu et al., 2018) proposes a two-part novel approach. The first part utilizes the underlying structure of the data space, discovered based on density data peaks, to train more robust classifiers. The second part involves using differential evolution to refine the location of newly classified data during the self-training process. "Newly labeled data" refers to unlabeled data labeled by the classifier during self-training, and "maximizing the location" ensures optimal balance in the attribute date values. (Livieris et al., 2018) Presents a new semi-supervised learning algorithm based on self-training and proposes an algorithm that automatically selects the best base learner relative to the number of the most confident predictions of unlabeled data. The proposed algorithm's performance is tested on various benchmark datasets regarding classification accuracy using commonly used simple learners. (Hyams, 2017b) examines the intuitive and flexible self-training approach as a semi-supervised approach for computer vision tasks. (Piroonsup & Sinthupinyo, 2018) propose a new method to determine the sufficiency of labeled data by applying a semi-supervised cluster technique to estimate the labeled data distribution over the training set and suggest two methods to improve the labeled dataset in the insufficient portion. The results show that the accuracy obtained from the final classifiers in clusters without labeled data is markedly lower than that obtained from clusters with labeled data. (J. Li & Zhu, 2019) introduce a new self-training method based on an optimum path forest, comprising three main parts:

1. They propose constructing an optimum path forest to discover the potential spatial structure of the feature space.
2. They use the structure to guide self-training methods to iteratively label unlabeled samples, which are then used to expand the labeled data.
3. A desirable classifier can be trained with the extended labeled data.

- **Co-training methods:** Co-training is a machine learning algorithm used when there is a small amount of labeled data and a significant amount of unlabeled data. It is a semi-supervised learning method with two viewpoints, assuming that every sample is described using two sets of various features, presenting different information. These two views are conditionally independent, and each is sufficient for classification. Co-training learns separate classifiers for each view using labeled samples. (Xing et al., 2018) introduce a solution to address the issue of class imbalance called multi-label co-training (MLCT). It interacts with confident labels of multi-label samples during the co-training process. MLCT implements a predictive reliability test to select samples and employs label-wise filtering to assign labels to the selected samples confidently. Experimental findings indicate that the suggested approach outperforms other similar co-training classifiers.

Generative model methods:

This procedure involves utilizing unlabeled data for more accurate evaluations. Various models have been introduced for semi-supervised learning. Generative models, such as mixed Gaussian distribution, the EM algorithm, Bayesian distribution, hidden Markov models, and the Baum-Welch algorithm (Kumar et al., 2016), are based on iterative approaches. In particular, (Rezende et al., 2016) have developed a new class of general-purpose models with a single-shot generalization capability, emulating an essential characteristic of human cognition. However, the proposed approach still has some limitations. It requires a reasonable amount of data to avoid overfitting.

Margin-based methods:

Supervised margin-based methods have proven to be successful techniques for classification. Many studies have been conducted to extend these methods to the domain of semi-supervised learning. For instance, (Kaneko, 2019) proposed a novel online multiclass classification algorithm based on the forecast margin for partial feedback settings. The suggested technique focused on the forecast margin and learning from complementary labels in online classification. Experimental results have demonstrated that the proposed algorithm significantly outperforms other methods in the same setting.

Graph-based methods:

Graph-based semi-supervised learning methods are rooted in graph theory. These methods define a graph where nodes represent labeled or unlabeled samples, and edges indicate the similarities between samples. Typically, these methods assume label evenness within the graph. Many graph-based methods aim to estimate a function on the graph. (Martineau et al.,

DOI: 10.6977/IJoSI.202403_8(1).0002
O. Stitini, S. Kaloun, etc./Int. J. Systematic Innovation, 8(1), 11-26 (2024)

2020) Present a practical scheme for optimizing the graph-matching problem in a classification context. They propose a representation based on a parametrized model graph and optimize the associated parameters to enhance classification accuracy. (Yang et al., 2016) suggest a unified framework that directly operates on multi-class problems without reducing them to binary tasks. This framework also enables practical feasibility for active learning in multi-class scenarios, which the one-vs-all strategy cannot achieve. (L. Wang et al., 2018) Designed an Adaptive Graph Guided Embedding (AG2E) approach for a semi-supervised multi-label learning scenario. AG2E leverages limited labeled data and unlabeled data to improve multi-label learning performance.

Semi-supervised boosting:

Boosting is a supervised learning method with numerous applications. The primary objective of boosting is to minimize marginal costs. Additionally, this method has been extended and developed for semi-supervised learning (Tanha, 2019).

3.2.2 Clustering approach

Clustering can serve as a means of summarizing the distribution of samples. It is often employed before the classification stage to reduce unnecessary details. Semi-supervised clustering approaches fall into the category of clustering methods that can be extended to handle partially labeled data or data with other outcome steps (sometimes referred to as supervised clustering methods). Numerous algorithms have been developed for semi-supervised clustering, including hierarchical, partitioning, density-based, grid-based, and model-based methods.

Hierarchical methods:

(Nakano et al., 2020) Proposed a method to enhance accuracy in multi-class classification tasks. The idea behind their approach is that in situations with many classes, traditional methods may need help to correctly classify new observations due to the sheer number of possibilities. To address this, the researchers suggest building specialized classifiers for classes that often result in common misclassifications. In other words, they propose constructing a chain of specialized classifiers to handle simpler subproblems.

Partitioning methods:

(Karimi et al., 2018) Propose a partnership between business and user reviews to forecast multi-label grouping and introduce a mix of k-means between business and user reviews. The effectiveness of machine learning algorithms heavily relies on the chosen data representations or attributes, with abundant and efficient representations leading to strong prediction outcomes. Some machine learning algorithms, like deep learning, can learn the representations mapping to outputs and the representations themselves. However, these algorithms require a significant volume of data to obtain usable representations, which is often unavailable in outlier mining. Nevertheless, this principle is directly adaptable to outlier detection. Unsupervised outlier detection techniques can extract richer representations from small datasets, also known as unsupervised feature engineering. This method has enhanced data expression and optimized supervised learning (Jedrzejowicz et al., 2018).

Density-based methods

(Gertrudes et al., 2019) suggests a semi-supervised self-training classification algorithm based on data density peaks and differential evolution.

3.2.3 Deep learning approach

As a subset of machine learning, deep learning is based on algorithms designed to model high-level abstract concepts in databases. Deep learning finds applications in various image classification tasks, such as object identification, image extraction, semantic segmentation, and gesture estimation. In this study, we aim to compare the differences between existing famous works on fake news detection using the same dataset and the results obtained from our new proposed approach in section 3.
(Karimi et al., 2018) Introduce an approach to combine information from multiple sources and distinguish between different degrees of fakeness. They propose a Multi-source, Multi-class Fake News Detection framework (MMFD). The proposed system combines automated extraction features, multi-source fusion, and fakeness detection into a coherent and interpretable model. Experimental results demonstrate the viability of the proposed framework, and extensive experiments are conducted to gain insights into its

workings.

4. Proposed approach

4.1 Aim of the study

Specifying the meaning and type of information required to classify an item as fake news is crucial. Moreover, specifying the form of fake news is beneficial, considering the various types ranging from low to high. Any news analysis must rely on a formal classification of incorrect information, including propaganda, satirical information, manipulation, and manufacturing. However, our primary interest lies not in developing a general classification procedure but in building an automatic algorithm capable of multi-classifying any news while providing a specific percentage for each form.

To achieve our proposed approach's goal, we encounter two key challenges:

1. Differentiating between various forms of fake news.
2. Implementing multi-class classification using unlabeled data.
   a. Estimating labels based on similarity to enhance and improve the self-training algorithm.
   b. Comparing the newly estimated labels using similarity with the new labels predicted by the voting majority.

4.2 Methodology and approach

This section outlines the comprehensive approach (refer to Fig. 4), which encompasses three main categories:

1. Input Phase: It comprises two significant steps: Data collection of news (labeled and unlabeled data) and the pre-processing using NLP.
2. Pre-processing phase: It contains three major steps: Vectorization, Recommender System, and Multi-class classification.
3. Output phase

4.2.1 Input phase:

In the Data collection phase, we gather both labeled and unlabeled news articles from various sources.Labeled data refers to news articles that have been manually classified into different categories, such as satire, propaganda, manufacturing, and manipulation. These labels serve as the ground truth for training and evaluating our models. Unlabeled data, on the other hand, consists of news articles that have not been classified into any specific category.

The pre-processing step involves preparing the collected news articles for further analysis and classification using Natural Language Processing (NLP) techniques. In this phase, we perform various operations to clean, normalize, and transform the text data, making it suitable for machine learning algorithms.

The pre-processing steps typically include:

- Text Cleaning: Removing irrelevant or noisy elements, such as HTML tags, special characters, punctuation, and numbers.
- Tokenization: Breaking down the text into individual words or tokens to facilitate further analysis.
- Stopword Removal: Eliminating common words (e.g., "the," "is," "a") that do not contribute much to the overall meaning of the text.
- Lowercasing: Converting all words to lowercase to ensure uniformity and avoid treating words with different cases as distinct.
- Lemmatization or Stemming: Reducing words to their base or root form to reduce the vocabulary size and improve text representation.
- Removing Rare Words: Eliminating words that occur very infrequently, as they may not contribute significantly to the overall meaning.

After the pre-processing steps, we have a clean and transformed dataset that can be used for feature extraction and subsequent classification. For feature extraction, we employ the Word2Vec word embedding technique, which converts words into dense vector representations. These vectors capture the semantic meaning of the words and their contextual relationships, enabling better text representation for classification tasks. Overall, the data collection and pre-processing phase plays a crucial role in preparing the input data for our multi-class semi-supervised approach, ensuring that the data is in a suitable format for further analysis and model training.

4.2.2. Pre-processing phase:

This phase involves three main steps: Vectorization, Recommender System, and Multi-class classification.

- Vectorization:
  In the Vectorization step, both labeled and unlabeled news data are transformed into numerical vectors. Each news article's text is converted into a vector representation. We utilize the widely adopted word embedding technique, Word2Vec, as it offers a common and effective way to represent textual vocabulary. Word2Vec can capture the semantic meaning of words in the context of a text, capturing textual and syntactic similarities, as well as the relationships between different words. By using Word2Vec, we can create meaningful and compact representations of the news articles, which will be used as input for our subsequent classification process.
- Recommender System
  In the Recommender System phase, we perform a pre-multi classification process consisting of two steps:
- **Similarity recommendation:** In this step, we predict the label for each unlabeled news text by calculating its similarity to the entire labeled dataset. We use a similarity measure to determine how closely each unlabeled text aligns with the labeled data. The text with the highest similarity score is considered the most similar to a labeled text, and its predicted label is assigned accordingly.
- **Majority voting recommendation:** To further refine our predictions, we train a model using the labeled data and then use this trained model to predict labels for the unlabeled data using a voting majority approach. This means that we apply multiple algorithms to make predictions for each unlabeled text, and the final label is determined by selecting the most frequently predicted value among these algorithms.
- **Multi-class classification phase**: we utilize the recommendations obtained from the previous steps, which include two label suggestions for each news article: one based on similarity and the other based on voting majority. News articles identified as similar to the labeled data, their corresponding recommended labels are incorporated into the labeled dataset, effectively creating a new pseudo dataset. This process enriches the labeled dataset with additional samples, enhancing the diversity and representativeness of the training data. However, for news articles deemed dissimilar to the labeled data, we employ the self-training algorithm as the third approach to predict their labels. The self-training algorithm utilizes the information from the labeled dataset to make predictions for the unlabeled data, enabling us to assign labels to these non-similar articles. By combining these different strategies, our multi-class classification approach aims to effectively handle diverse news articles, leveraging similarity-based recommendations and self-training predictions to improve the accuracy and comprehensiveness of the final classification results.

By employing the Recommender System, we aim to improve the self-training algorithm's effectiveness and obtain more accurate and reliable multi-class classification results for news articles.

4.2.3. Output phase:

In the final phase, the proposed approach provides the predicted classes of the given news text and the corresponding percentage of certainty for each class.

Self-training is a powerful learning method that effectively handles situations with limited labeled and abundant unlabeled data. However, it is often observed that the accuracy of applying the unlabeled data in addition to the labeled data is lower than using only the labeled data. One of the main reasons for this is the inadequacy of the labeled data to train the initial classifier in the self-training phase. An inefficient initial classifier can introduce mislabeled data, which is then utilized to train the final classifier, leading to a decline in the precision of the semi-supervised self-training classifier. To address this issue, we propose a novel approach to ensure the newly labeled data's reliability for training the final classifier.

This section presents a method for constructing a multi-class classification analysis model (Fig. 3). The process of the proposed methodology is divided into the following steps:

- Step 1- Similarity Phase: Each unlabeled news item is transformed into a vector using word2Vec. Then, we calculate the similarity

DOI: 10.6977/IJoSI.202403_8(1).0002
O. Stitini, S. Kaloun, etc./Int. J. Systematic Innovation, 8(1), 11-26 (2024)

percentage by employing the Cosine Similarity between each unlabeled item and the entire labeled dataset. The most similar item (with the maximum similarity value) is identified, indicating that this unlabeled item has a percentage of similarity in terms of having the same type as the labeled item it resembles.

- Step 2- Voting Majority Phase: Initially, the model is trained on the labeled data, and then we utilize this pre-trained model to predict labels for the unlabeled data using five classifiers. The label predicted most frequently among the five classifiers is assigned to the unlabeled news item as a second prediction, which we call the voting majority.
- Step 3- Pre-Multi Classification Phase: In this step, we have two predicted types for each news item: one based on similarity with its respective percentage of similarity and the other based on the voting majority.
- Step 4- Recommendation Phase: News items with the same predicted types (i.e., the predicted type found in the similarity phase matches the one from the voting majority phase) are added to the labeled dataset. This phase enhances confidence in the items we will add to the dataset.
- Step 5- Multi-class Classification Phase: In this final step, we apply the self-training algorithm to the data, performing multi-class classification further to improve the accuracy and reliability of the predictions.

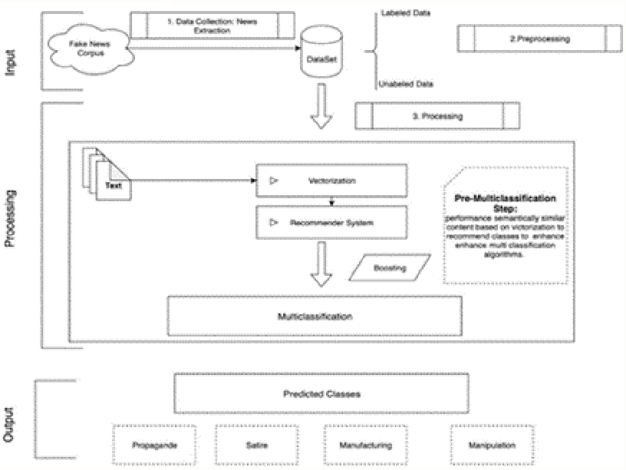

Fig. 2: The overall approach

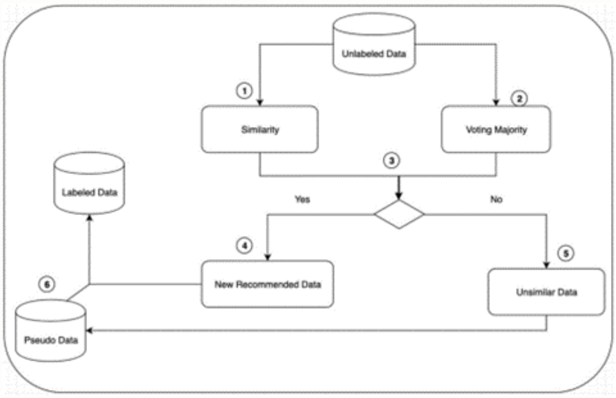

Fig. 3: Process of proposed approach

5. Experiments and results

5.1 Data collection process

The topic of fake news identification lacks standard benchmark datasets, primarily due to the term "fake news" encompassing various subcategories. Our study utilized two kinds of fake news identification datasets (Table 4). The first dataset contains three types of fake news: satire, propaganda, and manufacturing. The second dataset focuses on manipulation or bias form. We combined both datasets in a balanced manner. Each article in the corpus includes the title text and label. The corpus consists of 443 news articles for each label: satire, propaganda, manufacturing, and manipulation.

5.1.1 Dataset

The absence of manually labeled fake news datasets poses a significant challenge to the advancement of computationally expensive, text-based models that cover a wide range of topics. For our research purposes, we require a set of news articles that are directly classified into news types, such as satire, propaganda, manufacturing, and manipulation. We thoroughly searched for available datasets containing these news categories to address this issue. As a result, we found two datasets and combined them to create a comprehensive dataset with all four categories, facilitating multi-class classification.

Table 4: Sample fake news dataset.

References	Size	Date	Text
(Fact	12999	2017	This research analyzes

Checking)	rows		inaccurate news sources and truthful claims from politifact.com using tools from a previous EMNLP'17 paper
(Getting real about fake news)	38859 rows	2016-11-25	The study includes recent reports on non-fake news to emphasize the complexity of addressing inaccurate reporting and seeks better solutions than blacklists.

5.2. Data pre-processing

Pre-processing data is a typical first step that precedes training and evaluating data using machine learning algorithms. Ensuring the data is appropriately formatted and meaningful elements are integrated is crucial to achieving accurate and optimal outcomes. Our pre-processing of the data involved an iterative process divided into three main stages. Each incremental step corresponds to the models trained and evaluated on the pre-processed data at that stage. Moreover, each step builds upon the previous ones, with the second step including the first pre-processing stage and the third step including the first two pre-processing processes.

The first step is easy pre-processing, followed by the second, which involves removing all non-English phrases. Finally, the last step entails removing the end of the guardian posts, consistently including the exact phrase: "Share on x, y, z."

5.3. Experiment results

Our experimental results were obtained through a three-phase procedure. In the first phase, we calculated the similarity for each row in a small amount of labeled data and a large amount of unlabeled data, assigning a percentage of similarity. Table 5 shows the accuracy achieved after applying the new labels to the unlabeled data using similarity.

We trained the labeled data in the second phase as mentioned in Table 6 using different classifiers (Logistic regression, Naive Bayes, Linear SVM, and Decision Tree). We then used this model to predict the labels for the unlabeled data again, this time using a voting majority approach. We calculated the new percentage of recommendations by subtracting 100 from the similarity percentage found in the first phase. This phase is recommended for comparing the labels with the highest similarity percentage predicted in the first step with the labels predicted in this phase using a voting majority. Table 7 displays the results of this phase.

Moving on to the third phase, we combined the similar labels predicted in the preceding phases and added them to the labeled dataset, creating a new pseudo dataset. We applied the self-training algorithm to predict dissimilar labels. Table 8 shows the results obtained in this phase. Finally, we checked if the newly predicted labels matched those in the second phase.

Table 5: Experimental results for the first phase.

Different Models	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.76	0.76	0.76	0.76
Naive Bayes	0.44	0.69	0.44	0.39
Decision Tree	0.41	0.35	0.41	0.35
Linear SVM	0.73	0.74	0.73	0.73

Table 6: Experimental results for the second phase.

Different Models	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.93	0.93	0.93	0.91
Naive Bayes	0.71	0.73	0.71	0.66
Decision Tree	0.74	0.62	0.74	0.67
Linear SVM	0.94	0.94	0.94	0.93

Table 7: Experimental results for the third phase.

Different Models	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.97	0.97	0.97	0.97
Naive Bayes	0.64	0.64	0.47	0.55
Decision Tree	0.64	0.60	0.64	0.60
Linear SVM	0.95	0.95	0.95	0.94

Table 8: Performance comparison for fake news detection

Metric	(Collins et al.,2020).	(Wu et al.,2018)	(Stitini et al.,2022)	Our proposed approach
Accuracy	89%	33%	96%	97%
Precision	89%	-	96%	97%
Recall	89%	60%	96%	97%
F1-Score	89%	49%	96%	97%

6. Conclusion

In this paper, we introduce a novel multi-class semi-supervised approach for self-training, which is trained using a limited collection of classified data and an extensive amount of unlabeled data. Our innovative solution incorporates a similarity algorithm to enhance the self-training process, ensuring new expected labels are applied to the labeled data.

To evaluate the efficiency of the proposed semi-supervised method, we conducted tests on two benchmark datasets, measuring the classification precision using commonly available simple learners such as logistic regression, decision tree, naive Bayes, and linear SVM. The numerical findings validate the effectiveness and robustness of our approach. Consequently, our method contributes to developing more effective, reliable, and robust predictive models for multi-class fake news classification.

References

Kumar, S., West, R., & Leskovec, J. (2016). Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes. Proceedings of the 25th International Conference on World Wide Web.

Qian, F., Gong, C., Sharma, K., & Liu, Y. (2018). Neural User Response Generator: Fake News Detection with Collective User Intelligence. https://doi.org/10.24963/ijcai.2018/533

Collins, B., Hoang, D. T., Nguyen, N. T., & Hwang, D. (2020). Trends in combating fake news on social media – a survey. Journal of Information and Telecommunication, 1–20. https://doi.org/10.1080/24751839.2020.1847379

Vijayaraghavan, S. (2020, February 15). Fake News Detection with Different Models. arXiv.org. https://arxiv.org/abs/2003.04978

Li, Q., Zhang, Q., Si, L., & Liu, Y. (2019). Rumor detection on social media: datasets, methods, and opportunities. https://doi.org/10.18653/v1/d19-5008

Alzanin, S. M., & Azmi, A. M. (2018). Detecting rumors in social media: A survey. Procedia Computer Science, 142, 294–300. https://doi.org/10.1016/j.procs.2018.10.495

Wang, W. Y. (2017). “Liar, liar Pants on Fire”: a new benchmark dataset for fake news detection. https://doi.org/10.18653/v1/p17-2067

Wu, L., Rao, Y. J., Yu, H., Wang, Y., & Nazir, A. (2018). False information detection on social media via a hybrid deep model. In Lecture Notes in Computer Science (pp. 323–333). https://doi.org/10.1007/978-3-030-01159-8_31

Imran, M., Castillo, C. F., Diaz, F., & Vieweg, S. (2015). Processing social media messages in mass emergency. ACM Computing Surveys, 47(4), 1–38. https://doi.org/10.1145/2771588

Oumaima, S., Soulaimane, K., & Omar, B. (2020). Latest Trends in Recommender Systems applied in the medical domain. https://doi.org/10.1145/3386723.3387860

Tanha, J. (2019). A multiclass boosting algorithm to labeled and unlabeled data. International Journal of Machine Learning and Cybernetics, 10(12), 3647–3665.https://doi.org/10.1007/s13042-019-00951-4

Stitini, O., Kaloun, S., & Bencharef, O. (2022). Towards the detection of fake news on social networks contributing to the improvement of trust and transparency in recommendation systems: Trends and challenges. Information, 13(3), 128. https://doi.org/10.3390/info13030128

Stitini, O., Kaloun, S., & Bencharef, O. (2022a). Integrating contextual information into multi-class classification to improve the context-aware recommendation. Procedia Computer Science, 198, 311–316. https://doi.org/10.1016/j.procs.2021.12.246

Silva-Palacios, D., Ferri, C., & Ramírez-Quintana, M.J. (2017). Improving performance of multiclass classification by inducing class hierarchies. Procedia Computer Science, 108, 1692–1701. https://doi.org/10.1016/j.procs.2017.05.218

Rasool, T., Butt, W. H., Shaukat, A., & Akram, M. U. (2019b). Multi-Label Fake News Detection using Multi-layered Supervised Learning. https://doi.org/10.1145/3313991.3314008

Jedrzejowicz, J., Kostrzewski, R., Neumann, J., & Zakrzewska, M. (2018). Imbalanced data classification using MapReduce and Relief. Journal of Information and Telecommunication. https://doi.org/10.1080/24751839.2018.1440454

Forestier, G., & Wemmert, C. (2016). Semi-supervised learning using multiple clusterings with limited labeled data. Information Sciences, 361–362, 48–65. https://doi.org/10.1016/j.ins.2016.04.040

Karimi, H. R., Roy, P. C., Saba-Sadiya, S., & Tang, J. (2018). Multi-Source Multi-Class fake news detection. In International Conference on Computational Linguistics (pp. 1546–1557). https://www.aclweb.org/anthology/C18-1131.pdf

Kaliyar, R. K., Goswami, A., & Narang, P. (2019). Multiclass Fake News Detection using Ensemble Machine Learning. https://doi.org/10.1109/iacc48062.2019.8971579

Li, J., & Zhu, Q. (2019). Semi-Supervised Self-Training method based on an Optimum-Path forest. IEEE Access, 7, 36388–36399. https://doi.org/10.1109/access.2019.2903839

Wu, D., Shang, M., Wang, G., & Li, L. (2018). A self-training semi-supervised classification algorithm based on density peaks of data and differential evolution. https://doi.org/10.1109/icnsc.2018.8361359

Martineau, M., Raveaux, R., Conte, D., & Venturini, G. (2020). Learning error-correcting graph matching with a multiclass neural network. Pattern Recognition Letters, 134, 68–76. https://doi.org/10.1016/j.patrec.2018.03.031

Yang, P., Zhao, P., Hai, Z., Liu, W., Hoi, S. C. H., & Li, X. (2016). Efficient multi-class selective sampling on graphs. In Uncertainty in Artificial Intelligence (pp. 805–814). http://auai.org/uai2016/proceedings/papers/34.pdf

Kaneko, T. (2019, February 4). Online multiclass classification based on prediction margin for partial feedback. arXiv.org. https://arxiv.org/abs/1902.01056

Gertrudes, J. C., Zimek, A., Sander, J., & Campello, R. J. G. B. (2019). A unified view of density-based methods for semi-supervised clustering and classification. Data Mining and Knowledge Discovery, 33(6), 1894–1952. https://doi.org/10.1007/s10618-019-00651-1

Larriva-Novo, X., Sánchez-Zas, C., Villagrá, V. A., Vega-Barbas, M., & Rivera, D. (2020). An approach for the application of a dynamic Multi-Class classifier for network intrusion detection systems. Electronics, 9(11),1759. https://doi.org/10.3390/electronics9111759

Deepti Nikumbh, Anuradha Thakare (2023). A comprehensive review of fake news detection on social media: feature engineering, feature fusion, and future research directions, International Journal of Systematic Innovation, 7(6), 36-53. DOI: 10.6977/IJoSI.202306_7(6).0004

Livieris, I. E., Kanavos, A., Tampakas, V., & Pintelas, P. E. (2018). An Auto-Adjustable Semi-Supervised Self-Training Algorithm. Algorithms, 11(9), 139. https://doi.org/10.3390/a11090139

Hyams, G. (2017, September 30). Improved training for Self-Training by Confidence assessments. arXiv.org. https://arxiv.org/abs/1710.00209

Piroonsup, N., & Sinthupinyo, S. (2018). Analysis of training data using clustering to improve semi-supervised self-training. Knowledge Based Systems, 143, 65–80. https://doi.org/10.1016/j.knosys.2017.12.006

Xing, Y., Yu, G., Domeniconi, C., Wang, J., & Zhang, Z. (2018). Multi-Label Co-Training. https://doi.org/10.24963/ijcai.2018/400

Rezende, D. J., Mohamed, S., Danihelka, I., Gregor, K., & Wierstra, D. (2016). One-shot generalization in deep generative models. In International

Conference on Machine Learning (pp. 1521–1529). http://jmlr.org/proceedings/papers/v48/rezende16.pd f

Wang, L., Ding, Z., & Fu, Y. (2018). Adaptive Graph guided embedding for multi-label annotation. https://doi.org/10.24963/ijcai.2018/388

Nakano, F. K., Cerri, R., & Vens, C. (2020). Active learning for hierarchical multi-label classification. Data Mining and Knowledge Discovery, 34(5), 1496–1530. https://doi.org/10.1007/s10618-020-00704-w

Fact Checking. https://hrashkin.github.io/fact-check.html Accessed: 2021-02-24

Getting Real about Fake News. https://www.kaggle.com/datasets/mrisdal/fake-news . Accessed: 2021-02-24

Stitini, O., Kaloun, S., & Bencharef, O.: Investigating different similarity metrics used in various recommender systems types: scenario cases, Int. Arch.Photogramm. Remote Sens. Spatial Inf. Sci., XLVIII-4/W3-2022, 187–193, https://isprs-archives.copernicus.org/articles/XLVIII-4-W3-2022/187/2022/

AUTHOR BIOGRAPHIES

Oumaima Stitini. She is currently an Assistant Professor at the Cadi Ayyad University, ENS. She received his Ph.D. in Computer Science Engineering, especially in Robust optimization and implementation of recommendation systems based on artificial intelligence from the Faculty of Science and Technology in 2022.
His main research interests include artificial intelligence, recommender systems, and IoT all of these applied to different fields like medical, education, and entertainment.

Sara Qassimi holds a Ph.D. and an engineering degree in computer science from Cadi Ayyad University, Marrakesh, Morocco. She is currently an Assistant Professor in the Department of Computer Science at the Faculty of Sciences and Technics Guiliz, and is affiliated with the L2IS Laboratory at Cadi Ayyad University. Her research interests focus on artificial intelligence, recommender systems, context-aware systems, social interactions, and knowledge graphs.

Soulaimane Kaloun He is currently holding the position of a Permanent Associate Professor at the Faculty of Science and Technology located in Marrakech, Morocco. He earned his doctorate degree in Computer Science and is presently serving as a professor at the same institution. Moreover, he has also received an HDR in data science. Soulaimane's principal areas of research revolve around Big Data, machine learning, multiagent systems, and text-mining.

Omar Bencharef. He is currently a Permanent Professor at the Faculty of Science and Technology, Marrakech, Morocco. Omar received his Ph.D. in Computer Science and is currently a professor at the Faculty of Science and Technology, Marrakech, Morocco. He's also HDR in data science. His main research interests include artificial intelligence (AI), data science, machine learning, multiagent systems, and text-mining.

DOI: 10.6977/IJoSI.202403_8(1).0002
O. Stitini, S. Kaloun, etc./Int. J. Systematic Innovation, 8(1), 11-26 (2024)

Works	Semi-supervised learning approach						Clustering approach			Deep Learning approach		
	Self-training	Co-training	GMM	MM	GM	BM	PM	HM	DM	GrM	CNN	RNN
(Tanha,2019).						✓						
(Kaliyar et al., 2019)						✓					✓	
(Silva-Palacios et al., 2017).						✓		✓				
(Li et al., 2019)	✓											
(Wu et al., 2018)	✓											
(Martineau et al., 2020)			✓		✓						✓	
(Yang et al., 2016)					✓							
(Kaneko, 2019)							✓					
(Larriva-Nov o et al., 2020										✓		
(Gertrudes et al., 2019)									✓			
(Deepti Nikumbh, et al., 2023)		✓										
(Karimi et al., 2018)												✓
(Kaneko et al., 2019)				✓								

Note: GMM refers to Generative models methods, MM refers to Margin based methods, GM refers to Graph based methods, BM refers to Boosting methods, PM refers to Partitioning Methods, HM refers to Hierarchical Methods, DM refers to Density-based Methods, GrM refers to Grid-based Methods, CNN refers to Convolutional neural networks, RNN refers to Recurrent Neural Network.