

UNIVERSIDADE FEDERAL FLUMINENSE

JULLIANO TRINDADE PINTAS

# Crowd-based Feature Selection for Text Classification

NITERÓI

2020

UNIVERSIDADE FEDERAL FLUMINENSE

JULLIANO TRINDADE PINTAS

## Crowd-based Feature Selection for Text Classification

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Doutor em Computação. Área de concentração: Engenharia de Sistemas e Informação.

Orientador:

LEANDRO AUGUSTO FRATA FERNANDES

Co-orientador:

ANA CRISTINA BICHARRA GARCIA

NITERÓI

2020

Ficha catalográfica automática - SDC/BEE  
Gerada com informações fornecidas pelo autor

P659c    Pintas, Julliano Trindade  
          Crowd-based Feature Selection for Text Classification /  
          Julliano Trindade Pintas ; Leandro Augusto Frata Fernandes,  
          orientador ; Ana Cristina Bicharra Garcia, coorientadora.  
          Niterói, 2020.  
          73 f. : il.

          Tese (doutorado)-Universidade Federal Fluminense, Niterói,  
          2020.

          DOI: <http://dx.doi.org/10.22409/PGC.2020.d.12434696767>

          1. Seleção de Atributos. 2. Inteligência Coletiva. 3.  
          Redução de dimensionalidade. 4. Classificação de texto. 5.  
          Produção intelectual. I. Fernandes, Leandro Augusto Frata,  
          orientador. II. Garcia, Ana Cristina Bicharra, coorientadora.  
          III. Universidade Federal Fluminense. Instituto de  
          Computação. IV. Título.

CDD -

JULLIANO TRINDADE PINTAS

Crowd-based Feature Selection for Text Classification

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Doutor em Computação. Área de concentração: Engenharia de Sistemas e Informação.

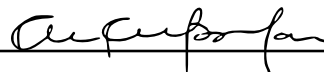
Aprovada em dezembro de 2020.

BANCA EXAMINADORA



---

Prof. Leandro Augusto Frata Fernandes - Orientador, UFF



---

Profa. Ana Cristina Bicharra Garcia - Co-orientadora,  
UNIRIO



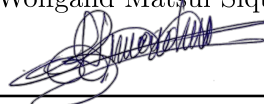
---

Prof. Luís Miguel Parreira e Correia, ULisboa



---

Prof. Sean Wolfgang Matsui Siqueira, UNIRIO



---

Prof. Flávia Cristina Bernardini, UFF

Niterói

2020

*Aos meus pais, Joandyr e Claudia, que me apoiaram  
incondicionalmente durante todas as etapas da minha formação.*

# Agradecimentos

Agradeço aos meus orientadores, Ana Cristina Bicharra Garcia e Leandro Augusto Frata Fernandes, por todo tempo dedicado e apoio durante o doutorado. O acompanhamento de perto de ambos foi essencial para que eu conseguisse avançar e desenvolver todas as atividades de pesquisa.

Aos professores e funcionários do Instituto de Computação da Universidade Federal Fluminense, por todo ensinamento e compromisso com os alunos.

Aos colegas de trabalho e gerentes da Petrobras que me apoiaram e permitiram conciliar as atividades do doutorado com as minhas demais atividades profissionais.

Por fim, agradeço a todos familiares e amigos que estiveram ao meu lado durante este período e que foram compreensivos em todos os momentos que precisei ficar focado nas atividades acadêmicas.

# Resumo

Os métodos de Seleção de Atributos (SA) aliviam problemas principais no desenvolvimento de modelos de classificação de textos, pois são utilizados para reduzir a dimensionalidade dos dados, melhorar a acurácia do modelo e reduzir o custo computacional. Por este motivo, os métodos de SA têm recebido muita atenção da comunidade de inteligência artificial nos últimos anos. Realizamos uma Revisão Sistemática de Literatura (RSL) abrangente que avaliou 1376 artigos únicos de periódicos e conferências publicados nos últimos oito anos (2013-2020). Após a triagem de resumos e análise de elegibilidade de texto completo, mapeamos 175 estudos sobre SA especificamente para classificação de textos. Nós identificamos que praticamente todos os métodos de SA mapeados possuem uma grande dependência do volume de dados rotulados de treinamento. No entanto, o conjunto de dados rotulados disponível pode ser limitado em muitas situações, o que pode degradar a eficácia desses métodos. Por esse motivo, investigamos o uso da inteligência coletiva no processo de SA com objetivo reduzir esta dependência. Nesta tese, propomos e avaliamos o método CrowdFS (Crowd-based Feature Selection) que se combina a avaliação de diferentes indivíduos para apoiar SA para classificação de textos. Para avaliar a eficácia do método proposto, realizamos um primeiro experimento com uma equipe de especialistas de uma empresa multinacional de energia e um segundo experimento utilizando uma plataforma aberta de tarefas (Appen) com participantes não especialistas. Nossa avaliação de resultados demonstrou que o CrowdFS resultou em uma acurácia equivalente aos métodos existentes, porém com uma dependência menor de volume de dados rotulados. Além disso, quando combinamos o método CrowdFS com métodos existentes, identificamos uma melhora na acurácia da classificação em relação ao uso dos métodos SA existentes de maneira isolada. Além de avaliar a eficácia do método proposto, discutimos nesta tese questões relevantes que identificamos sobre utilização de inteligência coletiva no processo de SA, como modos de agregação de respostas, número necessário de participantes, perfil dos participantes e critérios de qualidade.

**Palavras-chave:** Seleção de Atributos; Redução de dimensionalidade; Classificação de Textos; Inteligência Coletiva.

# Abstract

Feature Selection (FS) methods alleviate key problems in the development of text classification models as they are used to reduce the data dimensionality, improve the model accuracy, and reduce the computational cost. For this reason, FS methods have received a great deal of attention from the artificial intelligence community in recent years. We conducted a Systematic Literature Review (SLR) that assesses 1376 unique papers from journals and conferences published in the past eight years (2013-2020). After abstract screening and full-text eligibility analysis, we mapped 175 FS studies specifically for text classification. We identified that virtually all the mapped FS methods have a great dependence on the volume of labeled training data. However, the available labeled data set can be limited in many situations, which can degrade the effectiveness of these methods. For this reason, we investigated the use of collective intelligence in the FS process to reduce this dependence. In this thesis, we propose and evaluate the CrowdFS (Crowd-based Feature Selection) method that combines the evaluation of different individuals to support FS for text classification. To evaluate the effectiveness of the proposed method, we conducted a first experiment with a team of specialists from a multinational energy company and a second experiment using an open platform (Appen) with non-specialist participants. Our evaluation of results demonstrated that CrowdFS resulted in an accuracy equivalent to the existing methods, but with less dependence on the volume of labeled data. In addition, when we combined the CrowdFS method with existing methods, we identified an improvement in the accuracy of the classification in comparison to the use of existing FS methods in isolation. In addition to evaluating the effectiveness of the proposed method, we discussed in this thesis relevant issues that we identified about the use of collective intelligence in the FS process, such methods for aggregating responses, the required number of participants, profile of participants and quality criteria.

**Keywords:** Dimensionality reduction; Feature Selection; Text Classification; Collective Intelligence;



# List of Figures

2.1	Flowchart of the text classification process with state-of-the-art elements. The activities marked in gray are detailed in this section. Adapted from [126]. . . . .	5
3.1	PRISMA flow diagram for this Systematic Literature Review (SLR). . . . .	14
3.2	The four FS sub-tasks for text classification. . . . .	14
3.3	Number of studies by FS issue group over the years. The gray box indicates that data collected for 2020 may be preliminary since this review was updated in October 2020. . . . .	16
3.4	Proposed categorization schema for FS methods for text classification. Each vertical represents a different categorization perspective. We used this categorization scheme to map and analyze the 175 FS methods included in this review. . . . .	22
3.5	Flow diagram for each Feature Selection (FS) Strategy presenting the interaction between the FS activity and the model training activity. Each of these strategies is detailed in Section 3.3.1. . . . .	24
3.6	Number of FS studies by strategy over the years. The gray box indicates that data collected for 2020 may be preliminary since this review was updated in October 2020. . . . .	24
3.7	Classifiers that have been most often used to evaluate FS methods over the years. The gray box indicates that data collected for 2020 may be preliminary since this review was updated in October 2020. . . . .	40
3.8	Most used validation methods used to evaluate FS methods over the years. The gray box indicates that data collected for 2020 may be preliminary since this review was updated in October 2020. . . . .	41

3.9	Percentage of Filter Strategies vs Percentage of Wrapper Strategies over the years. The gray box indicates that data collected for 2020 may be preliminary since this review was updated in October 2020. . . . .	43
3.10	Number of FS studies by approach over the years. The gray box indicates that data collected for 2020 may be preliminary since this review was updated in October 2020. . . . .	45
3.11	Number of FS studies by type of classification over the years. The gray box indicates that data collected for 2020 may be preliminary since this review was updated in October 2020. . . . .	46
4.1	Crowd-based Feature Selection (CrowdFS) . . . . .	51
5.1	Demographics of experiment participants . . . . .	55
5.2	Average recall (macro and micro) by evaluation method . . . . .	58
5.3	Average precision (macro and micro) by evaluation method . . . . .	58
5.4	Macro averaged F-measure by evaluation method . . . . .	59
6.1	Features evaluation functionality developed in Appen and used for the second experiment. Experiment variation: 1 category per task (Group 1). . .	64
6.2	Features evaluation functionality developed in Appen and used for the second experiment. Experiment variation: 6 categories per task (Group 2). . .	65
6.3	Effectiveness comparison between unsupervised CrowdFS and the unsupervised baseline (TFIDF). Experiment variation: 1 category per task (Group 1). . . . .	66
6.4	Effectiveness comparison between unsupervised CrowdFS and the unsupervised baseline (TFIDF). Experiment variation: 6 categories per task (Group 2). . . . .	67
6.5	Effectiveness comparison between unsupervised CrowdFS and the supervised method (Chi-Square). Experiment variation: 6 categories per task (Group 2). . . . .	69

# List of Tables

2.1	Summary of advantages and disadvantages of text classification architectures. Adapted from Kowsari <i>et al.</i> [82]. . . . .	9
3.1	Number of FS studies for text classification by issue group and categorized according to the proposed categorization schema. All included studies were published between January/2013 and October/2020. Note: We use some abbreviations to simplify this table. TS refers to Two-Stages, and ML refers to Machine Learning. . . . .	23
3.2	Filter methods for measure relevance. . . . .	25
3.3	Filter methods for subset search, globalization, and ensemble issues. . . . .	26
3.4	FS studies using binary target text datasets. . . . .	33
3.5	FS studies grouped by labeled data dependence and year of publication. . . . .	36
3.6	Text representation models to evaluate FS methods. . . . .	38
3.7	Most commonly used public datasets to evaluate FS methods. . . . .	39
3.8	Most used language of text corpora in datasets to evaluate FS methods. . . . .	39
3.9	Classifiers that are most often used to evaluate FS methods. . . . .	40
3.10	Number of classifiers used to evaluate FS methods. . . . .	40
3.11	Validation methods used in experiments. . . . .	41
3.12	Statistical significance tests used in studies to reject or not the null hypothesis. . . . .	42
3.13	Filter Strategies versus Wrapper strategies over the years. . . . .	43
3.14	Number of supervised, unsupervised and semi-supervised studies over the years. . . . .	46
3.15	Age and size of most used datasets in experiments. . . . .	47
3.16	Historical trends in the usage of content languages for websites [199]. . . . .	47

---

3.17	Statistical significance tests used in studies to reject or not the null hypothesis.	48
3.18	Statistical significance tests used in studies to reject or not the null hypothesis (Conference Studies versus Journal Studies).	49
5.1	Number of training and testing documents of selected labels.	54
5.2	The intersection between the average evaluation of the specialists compared with an average evaluation of the other participants.	56
5.3	Comparison of CHI and BOS recall and precision results.	60
5.4	Percentage Increase in Precision, Recall and F-measure metrics.	60
5.5	P-Value for each group of results.	61
6.1	Number of participants, duration, and costs for the second experiment.	65
6.2	Effectiveness comparison between unsupervised CrowdFS and the unsupervised baseline (TFIDF).	67

# Table of Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Background</b>	<b>4</b>
2.1	Representation Models for Textual Data . . . . .	5
	<i>N</i> -gram based Representations . . . . .	5
	Word Embedding Representations . . . . .	6
2.2	Dimensionality Reduction . . . . .	7
	Feature Projection . . . . .	7
	Feature Selection . . . . .	8
2.3	Text Classification Architectures . . . . .	8
<b>3</b>	<b>Feature Selection Methods for Text Classification: A Systematic Literature Review</b>	<b>10</b>
3.1	Systematic Literature Review Protocol and Execution . . . . .	12
	3.1.1 Research Questions and Search Strategy . . . . .	12
	3.1.2 Conducting the Review . . . . .	13
3.2	Feature Selection Issues for Text Classification . . . . .	14
	3.2.1 Issues About Measure Feature Relevance . . . . .	16
	Unbalanced or Skewed Datasets . . . . .	16
	Rare Terms/Features . . . . .	17
	Redundancy Exclusion . . . . .	17
	Data Sparsity . . . . .	17

---

	Position/Location Inside the Text . . . . .	18
3.2.2	Issues About Subset Search . . . . .	18
	Search Strategy/Search Efficiency and Effectiveness . . . . .	18
	Redundancy . . . . .	19
	Feature Selection and Hyperparameter Optimization . . . . .	19
3.2.3	Issues About Globalization . . . . .	19
	Classes/Labels Representativeness on Final Feature Set . . . . .	20
	Class/Label Specific Features . . . . .	20
3.2.4	Issues About Ensemble . . . . .	21
3.3	Feature Selection Methods for Text Classification . . . . .	21
3.3.1	Categorization by Strategy . . . . .	23
	Filter Strategy . . . . .	24
	Wrapper Strategy . . . . .	26
	Embedded Strategy . . . . .	27
	Two-stages Hybrid Strategy . . . . .	27
	Two-stages Pure Strategy . . . . .	28
3.3.2	Categorization by Approach . . . . .	28
	Statistic-Based Approaches . . . . .	29
	Metaheuristic Approaches . . . . .	30
	Machine Learning-Based Approaches . . . . .	31
	Semantic-Based Approaches . . . . .	31
	Rule-Based Approaches . . . . .	32
	Linguistics Approaches . . . . .	32
3.3.3	Categorization by Target . . . . .	32
	Binary Text Classification . . . . .	33
	Multiclass Text Classification . . . . .	33

---

	Multi-Label Text Classification . . . . .	33
	Hierarchical Text Classification . . . . .	34
	Ordinal Text Classification . . . . .	34
3.3.4	Categorization by Labeled Data Dependence . . . . .	35
	Supervised Methods . . . . .	35
	Unsupervised Methods . . . . .	35
	Semi-Supervised Methods . . . . .	36
3.4	Experiment Settings Analysis . . . . .	36
3.4.1	Text Representation Used in Experiments . . . . .	37
3.4.2	Datasets Used in Experiments . . . . .	37
	Public Datasets . . . . .	37
	Language of Text Corpora in Datasets . . . . .	38
3.4.3	Classification Algorithms Used in Experiments . . . . .	39
3.4.4	Validation Settings Used in Experiments . . . . .	39
	Validation Method . . . . .	40
	Statistical Significance Test . . . . .	42
3.5	Discussion . . . . .	42
3.5.1	Filter has Been the Feature Selection Dominant Strategy for Text Classification, but a Change is Coming . . . . .	42
3.5.2	Metaheuristic Approach is the Trend . . . . .	44
3.5.3	Multiclass Classifiers are Still Dominant . . . . .	45
3.5.4	Supervised Versus Unsupervised Feature Selection Methods . . . . .	46
3.5.5	Recent Researches Still Over Old Public Datasets: The Need for New Benchmarks . . . . .	47
3.5.6	The English Language Dominance . . . . .	47
3.5.7	Feature Selection is Already a Mature Field Allowing Statistical Evaluations . . . . .	48

---

<b>4</b>	<b>Crowd-based Feature Selection Method for Text Classification (CrowdFS)</b>	<b>50</b>
<b>5</b>	<b>First Experiment (Supervised CrowdFS)</b>	<b>53</b>
5.1	Experiment Description . . . . .	53
5.2	Data Analysis . . . . .	55
5.3	Comparison of the average evaluation of experts and other participants . .	55
5.4	Comparison of the collaborative approach with the automatic approach . .	56
5.4.1	Analysis considering whole training set . . . . .	57
5.4.2	Analysis simulating small training sets . . . . .	59
5.5	Discussion about experimental results . . . . .	61
<b>6</b>	<b>Second Experiment (Unsupervised CrowdFS)</b>	<b>62</b>
6.1	Experiment Description . . . . .	63
6.2	Results Analysis . . . . .	65
6.3	Additional Analysis . . . . .	68
<b>7</b>	<b>Conclusion</b>	<b>70</b>
7.1	Feature Annotation vs Document Annotation . . . . .	70
7.2	Time and cost to use crowd platforms . . . . .	71
7.3	Quality criteria for responses . . . . .	72
7.4	Number of responses and participants . . . . .	73
	<b>References</b>	<b>74</b>
	Appendix A. List of Acronyms . . . . .	92



# Chapter 1

## Introduction

Automated text classifiers can be used to handle several real-world problems, such as spam filtering, sentiment analysis, and news classification [126, 30, 85]. Texts are usually represented by a high-dimensional and sparse document-term matrix in a space having the dimensionality of the size of the vocabulary containing word frequency counts. The high dimensionality can cause some problems, such as the curse of dimensionality and model overfitting. Feature Selection (FS) can be used to reduce dimensionality, remove irrelevant data, and increase the learning accuracy. FS is the process of automatically or manually select the features which contribute most to the classification of a given text. In text classification problems, the feature is usually some representation of a subset of words. A significant subset of features extracted from text corpora may not be relevant for the text classification task. These non-relevant features can either deteriorate the efficiency and accuracy of the classification models [86]. For this reason, FS for text classification became a popular research topic in artificial intelligence and data mining conferences and journals.

Despite receiving great attention from the text classification community, only a few literature surveys address FS focusing on text classification and the ones available are either a superficial analysis or present a very small set of work in the subject. To obtain an updated and comprehensive overview of the state of the art to identify the main open issues, we conducted a Systematic Literature Review (SLR) that assesses 1376 unique papers from journals and conferences published in the past eight years (2013–2020). After abstract screening and full-text eligibility analysis, 175 studies were included in our SLR. Our contribution in this SLR is twofold. We have considered several aspects of each proposed method and mapped them into a new categorization schema. Additionally, we mapped the main characteristics of the experiments, identifying which datasets, languages,

machine learning algorithms, and validation methods have been used to evaluate new and existing techniques. By following the SLR protocol, we allow the replication of our revision process and minimize the chances of bias while classifying the included studies. By mapping issues and experiment settings, our SLR helps researches to develop and position new studies with respect to the existing literature.

With our review, we identified that virtually all the mapped FS methods have a great dependence on the volume of labeled training data. However, the available labeled data set can be limited in many situations, which can degrade the effectiveness of these methods. A number of studies have already proven that aggregating the judgment of several individuals may result in estimates that are close to the real value in different domains, a phenomenon of collective intelligence known as wisdom of the crowds (WoC) [182]. The popularization of crowd-sourcing platforms and initiatives such as Amazon Mechanical Turk and Appen allow for an easy access to WoC. For this reason, this thesis introduces an approach based on collective intelligence to support FS for text classification. The central thesis statement of this research is presented as follows:

*The collective intelligence, through the aggregation of the judgments of several individuals about the relevance of features, can be combined to existing FS methods to improve the accuracy of resulting text classification model especially when the set of labeled training data is reduced.*

Two central issues are related to this thesis statement:

- **Improve Supervised FS Methods** - Apply collective intelligence to improve FS supervised methods that are dependent from labeled data.
- **Improve Unsupervised FS Methods** - Apply collective intelligence to improve FS unsupervised methods that are independent from labeled data.

I propose the CrowdFS (Crowd-based Feature Selection) method composed of two filter FS stages to explore these two central issues. In the first stage, existing FS methods can be used to perform the first filter of features. The second stage uses a collaborative evaluation to achieve the final feature selection. CrowdFS can be used as a hybrid FS method (Supervised + Unsupervised) using a supervised method in the first stage. However, it can also be used as a fully unsupervised FS method using an unsupervised FS method in the first stage. We performed two different experiments to evaluate the CrowdFS addressing each central issue. In the first experiment, we evaluated the CrowdFS using a supervised method, and the second experiment assessed the CrowdFS in a fully

unsupervised configuration. Performing two different experiments also allowed us to evaluate the CrowdFS in different settings and contexts.

We conducted the first experiment using the popular Chi-square method in the automatic stage of the CrowdFS. The collaborative stage of CrowdFS was performed out inside a Brazilian multinational energy company, where the participants were 29 of their employees. This quantitative experiment demonstrated this approach’s feasibility, resulting in better accuracy and coverage metrics compared with a popular automatic feature selection method in scenarios that are available with a small amount of labeled data.

Based on the analysis and discussion of the first experiment’s results, we designed a second experiment with a different configuration to evaluate the CrowdFS method. We conducted the second experiment using an open platform (Appen) available to anyone with access to the internet. In this second experiment, we also used a more recent dataset, had a larger number of participants (224 participants), evaluated a more extensive set of features in the collaborative stage (400 features). In the first experiment, we used a supervised method (Chi-square) in the early stage of CrowdFS. In this second experiment, we used an unsupervised method (TFIDF) for this purpose. In this way, we were able to evaluate the CrowdFS in a fully unsupervised configuration. In this second experiment, the CrowdFS method in unsupervised mode showed a relevant improvement in accuracy compared to the unsupervised baseline method. Additionally, the CrowdFS in unsupervised mode obtained accuracy similar to the supervised method when using a small amount of labeled data. This dissertation presents some original results that include:

- A general approach for crowd-based feature selection for text classification.
- A number of experimental evidences supporting that the proposed approach is feasible and can improve the accuracy of text classification models.
- The identification and discussion of challenges, issues, and future work based on the experiments’ results.
- A new categorization schema for FS methods for text classification created based on a comprehensive and recent literature review.

Chapter 2 presents background information about text classification and dimensionality reduction. Chapter 3 presents our systematic literature review that mapped 175 papers. Chapter 4 presents the proposed Crowd-based Feature Selection Method for Text Classification (CrowdFS). Chapter 5 and Chapter 6 presents the description and result analysis for first and second quantitative experiments. The challenges, issues, and future work related to this new FS approach are raised and discussed in Chapter 7.

# Chapter 2

## Background

Text classification is the problem to determine which class(es) a given document belongs to [120]. The classification problem can be divided into three main sub-types: binary, multiclass and multilabel. If only two classes are predefined, the problem is called as a binary classification problem. If three or more classes are defined, and each document can only be associated with one of these classes, it is known as a multiclass classification problem. Finally, if each document can be simultaneously associated with two or more classes (or labels), it is defined as a multilabel classification problem.

Apart from manual classification and hand-crafted rules, there is a third approach to text classification, namely, machine learning-based text classification [120]. In machine learning, the set of rules or, more generally, the decision criterion of the text classifier, is learned automatically from training data. This background section aims to present the state-of-the-art elements and activities to construct these classification models. Understanding the complete process for creating text classification models is essential to understand the applicability of FS in this process and its interaction with the other activities. Currently, developing models for text classification is a sophisticated process involving not only the training of models, but also numerous additional procedures, e.g., data pre-processing, transformation, and dimensionality reduction [126]. The activity flow including the main tasks for constructing text classification models is represented in Fig. 2.1.

We do not aim to present all the methods available for each activity exhaustively, as our study focuses on the FS activity. However, we will present some widely used techniques to explain the issues of each activity and their relationship with FS activity. To obtain a breakdown of the methods available for each stage, we recommend the recent

reviews on state-of-art elements and algorithms for text classification [126, 82].

This background section is organized into three parts, each detailing one of the main subactivities of the text classification pipeline. Section 2.1 introduces the key elements for feature extraction and weighting. Section 2.2 introduces the main concepts on dimensionality reduction, dividing into FS and feature projection. Finally, Section 2.3 presents learning algorithms for text classification.

## 2.1 Representation Models for Textual Data

Once you have labeled documents, the first step to construct a classification model is to extract features from text corpus. Different models of feature representation and weighting can be used for text classification and each representation model has advantages and disadvantages that must be considered, as discussed next.

***N*-gram based Representations** *N*-gram is a set of *N* words which occurs “in that order” in a text set [82]. The simplest and most widely used *N*-gram model is the BoW in which the  $N = 1$  (called 1-gram or uni-gram model). In this model, each feature corresponds a unique word in the text. However, the *N*-gram model can also be applied with *N* values greater than 1. For example, in the 2-gram model each feature corresponds to two consecutive words. *N*-gram models with  $N > 1$  could detect more information in comparison to 1-gram [82] because with  $N = 1$  the word order information is disregarded while in 2-gram or higher models part of the word order information is captured.

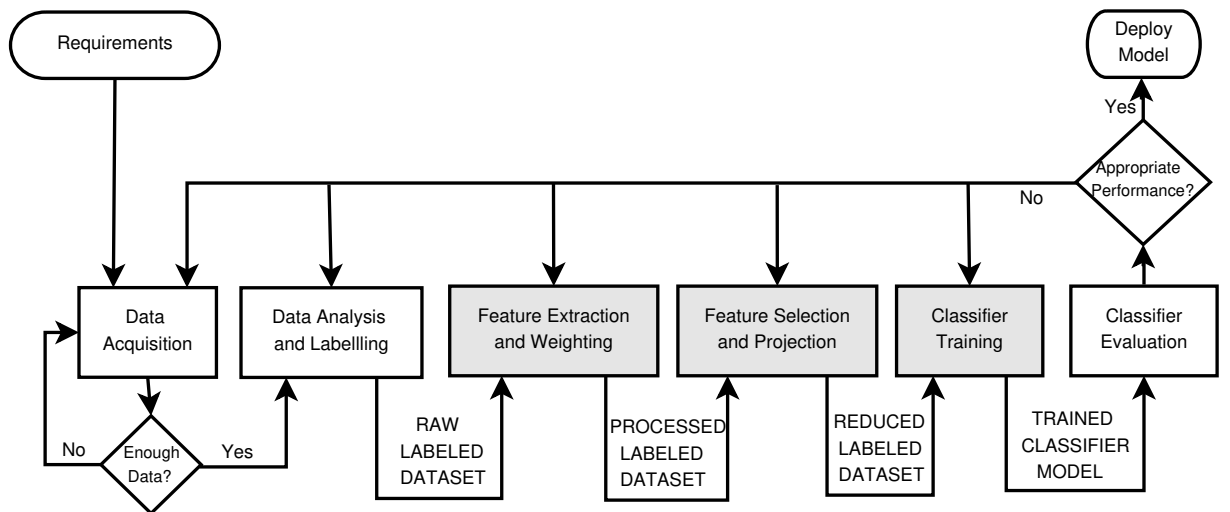


Figure 2.1: Flowchart of the text classification process with state-of-the-art elements. The activities marked in gray are detailed in this section. Adapted from [126].

In the  $N$ -gram model, each feature (a word or set of words) receives a value/weight for each document in the corpus. This value is usually calculated based on the frequency of that word (or set of words) in each document. The simplest is precisely the frequency of the word (or set of words) in the document, known as Term Frequency (TF). However, other weighting methods may be used. The most well-known and widely used method is the Term Frequency-Inverse Document Frequency (TF-IDF). In this method, the Inverse Document Frequency (IDF) is used in conjunction with TF in order to reduce the effect of implicitly common words in the corpus [82].

The main advantage of  $N$ -gram based models is their simplicity of both implementation and understanding. The main limitation is the fact that it fails to capture the similarity or semantics of the words. Additionally, in 1-gram models, information about the proximity of words within the text is not captured.

**Word Embedding Representations** The  $N$ -gram model is usually chosen to represent text in machine learning activities due to its simplicity, robustness and the observation that simple models trained on huge amounts of data outperform complex systems trained on less data [123]. However, recall that  $N$ -gram models don't measure the semantic similarity of the words becoming a limiting factor for some types of machine learning tasks [123]. Thus, many researchers have been looking for representation models that capture the syntactic or semantic similarity of words [123, 124, 82].

Unlike  $N$ -gram models that represent each word (or set of words) by a single value/weight per document, word embedding models represent each word (or set of words) by a  $N$ -dimension vector of real numbers [82]. The idea behind word embedding models is that similar words have vectors with close values. In this way, the level of syntactic or semantic similarity between words can be measured based on the distance of their vectors. Different techniques for estimating word vectors have been proposed, as Word2Vec [123], Glove [141] and FastText [26].

Despite having advantages of achieving a better representation of word similarity, word embedding models usually need very large corpora for training to present an acceptable vector for each word [157]. For this reason, in problems with small datasets, researchers prefer to use word embedding vectors that have been pre-trained on other large text corpora such as Google News with about 100 billion words [157]. However, pre-trained word embeddings may not include new words or domain-specific terms [82]. Therefore, the performance in the use of word embedding depends on the size of the corpus studied

and the availability of pre-trained word vectors appropriate to the context and language studied. These factors are relevant when choosing the best text representation model for each machine learning specific problem.

## 2.2 Dimensionality Reduction

As shown in Section 2.1, the main representation models used for text classification result in high-dimensional vectors. High dimensionality can cause some problems, such as the curse of dimensionality and model overfitting. For this reason, many researchers use dimensionality reduction techniques to produce smaller feature spaces [82]. According to Mironczuk and Protasiewicz [126], dimensionality reduction techniques can be organized into three groups: feature selection, feature projection, and instance selection. While the first two types of methods aim to reduce the dimensionality of the feature space, the third aims to reduce the number of instances used for training. In this section, we focus on feature selection and feature projection methods.

In feature selection methods, the resulting feature set is a subset of the initial feature set. On the other hand, the feature projection results in a new group of features mapped from the original features. Both methods can be used in isolation or combined to reduce dimensionality. Each method has its advantages and disadvantages, which will be detailed next.

**Feature Projection** Feature projection methods, also known as feature transformation methods, project the existing features onto different dimensions [126]. The main disadvantage of feature projection is the possible loss of meaning behind each original feature when data is transformed from the original dimensions into new dimensions. Depending on the projection method, there may not be a meaningful relationship between the projected and original dimensions. Feature projection have been studied less than feature selection methods for text classification [126]. Among the most popular feature projection methods are:

- Principal component analysis (PCA) [209]
- Linear Discriminant Analysis (LDA) [20]
- Latent Semantic Analysis (LSA) [94]
- Autoencoder [83]

**Feature Selection** Feature Selection (FS) methods are usually classified into three categories: filter, wrapper, and embedded [85]. This categorization is based on the FS strategy regarding how the FS integrates into the learning activity. Filter methods are executed as a previous step and are independent of the learning activity. Wrapper methods, on the other hand, encapsulate the predictor (i.e., the classifier) and utilize the performance of the predictor to assess the relevance of features or search the most relevant subset of features. Finally, embedded methods include FS as part of the training process.

The main advantage of selecting features is that the resulting feature set is a subset of the original set. This is an important point for text classification, as each feature usually represents a word or set of words. According to the survey work carried out by Mironczuk and Protasiewicz [126], feature selection is the most researched dimensionality reduction technique for text classification. In our SLR presented in next Chapter, we focus specifically on feature selection studies for text classification.

## 2.3 Text Classification Architectures

Over the years, different types of algorithms have been developed for the task of text classification [82]. These algorithms can be divided into two main groups: traditional machine learning and deep learning. Some traditional algorithms, like Support Vector Machines (SVM), Naive Bayes (NB) and  $k$ -Nearest Neighbors (KNN), are widely studied for the text classification problem and are still commonly used by the scientific community [82]. However, architectures based on deep learning like Convolutional Neural Network (CNN), Deep Belief Network (DBN), and Hierarchical Attention Network (HAN) are increasingly being researched for text classification [82]. Despite having the potential to achieve excellent results in some situations, deep learning architectures have some limitations and disadvantages. Table 2.1 compares deep learning and traditional architecture for text classification.

Table 2.1 shows that each text classification architecture has advantages and disadvantages. Thus, each specific situation must be analyzed before choosing between using deep learning or traditional architecture for text classification. Two central points in this choice are data volume and the need to have model interpretability. Deep learning usually requires much more data than traditional machine learning algorithms and not facilitate a comprehensive theoretical understanding of learning [82]. Therefore, if the volume of data available is small or there is a need for the interpretability of the model, the traditional



Table 2.1: Summary of advantages and disadvantages of text classification architectures. Adapted from Kowsari *et al.* [82].

Architecture Domain	Advantages	Disadvantages
Traditional	<ul style="list-style-type: none"> <li>• Requires a smaller amount of data</li> <li>• Models are less computationally expensive to train</li> <li>• Models have simpler interpretability</li> </ul>	<ul style="list-style-type: none"> <li>• Increase the need for feature engineering</li> <li>• Other model-specific disadvantages</li> </ul>
Deep Learning	<ul style="list-style-type: none"> <li>• Reduces the need for feature engineering</li> <li>• Can deal with complex input-output mappings</li> <li>• Parallel processing capability</li> </ul>	<ul style="list-style-type: none"> <li>• Requires a large amount of data</li> <li>• Is extremely computationally expensive to train</li> <li>• Complex model interpretability</li> </ul>

architecture will probably be more suitable.

The feature selection activity, which is the focus of this thesis, can be useful in traditional architecture and deep learning for text classification. As traditional architecture is more dependent on feature engineering activities, the selection of features has an important role in improving the classification model's accuracy. As the deep learning architecture has less dependence on feature engineering, feature selection tends to have less impact on the accuracy of the model. However, deep learning architectures are usually quite expensive to train. For this reason, feature selection may have an important utility for the deep learning architecture to reduce the computational cost.

## Chapter 3

# Feature Selection Methods for Text Classification: A Systematic Literature Review

As presented in the Chapter 2, dimensionality reduction plays a key role in text classification procedures. According to the survey work carried out by Mironczuk and Protasiewicz [126], feature selection is the most researched dimensionality reduction technique for text classification. In our SLR, we focus specifically on feature selection studies for text classification<sup>1</sup>. Some general reviews about FS are available. Chandrashekar and Sahin [30] and Kumar [85] provide a general introduction to FS methods and classify them into the filter, wrapper, and embedded categories. Pereira et al. [142] give a comprehensive survey and novel categorization of the FS techniques focusing on multi-label classification. However, these surveys did not consider in their analyses the different methods to handle the high dimensionality of the feature space, the different text representation formats such as bag of words and word embedding, and the power of the features' semantics for choosing the most efficient set of features.

FS methods have received a great deal of attention from the text classification community due to their strength in improving retrieval recall and computational efficiency [86]. However important, there are only a few literature surveys [86, 167, 40] that include them focusing on text classification. The ones available are either a superficial analysis or present a very small set of work in the subject. Kumbhar and Mali [86] and Shah and Patel [167] are more introductory studies, and both surveys don't focus only on FS methods. Besides to FS, Kumbhar and Mali [86] address feature extraction methods and

---

<sup>1</sup>The SLR presented in this chapter was submitted and accepted for publication in the journal Artificial Intelligence Review. URL: <https://www.springer.com/journal/10462>

Shah and Patel [167] address algorithms for text classification. For the best of our knowledge, there is only one review work focused exclusively on FS for text classification [40]. Although Deng et al. [40] provide a good overview of the subject, a limited proportion of published papers about FS for text classification have been included (28 studies). Among these, only fourteen were published in the last ten years, and six were published in the last five years. Besides, no clear criteria for inclusion or exclusion of the selected articles were defined. The study selection was made from other FS reviews that are not specific to text classification.

Our literature review expands existing surveys on FS methods, including up-to-date researches and providing a thorough analysis of FS methods considering the text classification task. The contribution of our literature survey lays on:

- Including a more significant number of papers covered (175 studies) resulting from a more comprehensive review in the theme;
- Bringing more up-to-date researches, including studies from 2013 to 2020;
- Proving a reproducible review according to an established literature review protocol;
- Providing a new research categorization for understanding the FS methods area;
- Providing a description of the experimental settings carried by the 175 reviewed studies; and
- Last but not least, we classified all 175 papers retrieved in our study according to our categorization scheme.

This chapter is organized as follows: The protocol of our SLR, which includes the research questions and inclusion/exclusion criteria for selecting the studies from the literature, is detailed in Section 3.1. Section 3.2 summarizes the issues addressed in the included studies. In Section 3.3, we cover all of the included studies by organizing them into a new categorization scheme specific to FS methods for text classification. The categorization schema proposed in this paper provides a simplified way to organize the actual methods as well as positioning new studies about FS for text categorization. The mapping of the included studies into this categorization schema allows us to identify which are the issues/topics that already have a significant number of studies and which ones have been less explored (possibly research gaps). In Section 3.4, we survey the experiment settings used to evaluate the proposed methods. We believe that the mapping of existing studies and their experiment settings would help researchers to position and develop new studies about FS for text classification.

## 3.1 Systematic Literature Review Protocol and Execution

The purpose of our review is to collect, organize in categories, and provide a comprehensive and recent review of FS methods for text classification. We decided to conduct a SLR to use a reproducible methodology and define explicit eligibility criteria. We aim to minimize the review bias and attempt to identify all studies that are related to our research questions.

There are several guidelines available to conduct SLRs, being Cochrane reviews protocol one of the most common in the health domain [64]. Based on Cochrane reviews protocol and other methods available in the literature, Kitchenham [80] proposed a protocol focused on software engineering. The SLR reported in this Chapter follows the Kitchenham's procedures for SLR.

We have performed a SLR in three databases: (1) IEEE Xplore Digital Library, (2) ACM Digital Library, and (3) Science Direct. Our SLR protocol includes the following steps: (i) the elaboration of research questions; (ii) the definition of search strategy; (iii) high-level paper selection and classification; and (iv) detailed review of selected papers. The searches were conducted using both title and abstract. It returned a total of 1376 unique papers from journals and conference considering the past eight years (2013–2020). After abstract screening and full-text eligibility analysis, 175 studies were included in our SLR.

### 3.1.1 Research Questions and Search Strategy

The purpose of this SLR is to find primary studies using an unbiased search strategy to answer the following research questions:

1. What are the main issues/problems that are being addressed by FS studies in text categorization task?
2. What are the different categories of methods that have been proposed?
3. What are the settings used to analyze and compare FS methods in experiments from the text categorization domain? For example: Text representation, Datasets, classifier algorithms and validation settings.

Preliminary searches were performed to assess the volume of potentially relevant studies. We identified that the query returned a small number of studies when applied only to

the studies' title. Searches using full text returned an impractical volume of non-relevant studies (dozens of thousands) because the searched terms are widespread in artificial intelligence literature. Therefore we decide to perform the search using title and abstract. Additionally, in our preliminary searches, we identified the words' main variants on the concepts we are looking for. Based on that, we construct our query string:

```
(Feature OR Features OR Variable OR Variables OR Attribute OR Attributes)
AND (Selection OR Select OR Selecting OR selected) AND (Text OR Texts OR
    Document OR Documents) AND (Categorization OR Classification OR
    Categorize OR Classify OR Categorizing OR Classifying OR Classifier)
```

### 3.1.2 Conducting the Review

Study selection refers to the assessment of retrieved papers. For this, we defined inclusion and exclusion criteria. The first exclusion criteria specified was based on practical issues (i.e., language and date of publication). This SLR considered papers published in English and between the years 2013 and 2020. The year restriction was established considering a large number of included studies in this period. The study selection activity was executed in two steps: (1) title and abstract screening; and (2) full text screening. We performed both steps manually.

In the first screening phase, papers were included only if they contain, either in the title or in the abstract, descriptions related to *Feature Selection* and *Classification Tasks* topics in *Text Domain*.

After the first screening step, full texts were retrieved and analyzed individually. At this point, the aim was to ensure that only those studies that are related to the subject considered in this review and that are related to our research questions would be selected. The following are the main reasons for studies exclusion after the full-text analysis: (1) The study does not focus exclusively on FS (70 studies). (2) The study does not evaluate the FS method using text datasets (33 studies). (3) The study does not evaluate the classification task's method (6 studies).

The SLR reported in this paper was conducted in October 2020. Included papers reached the number of 175 studies. Among these studies, 71 (40.57% of total) were retrieved from ACM Digital Library, 71 (40.57% of total) comes from IEEE Xplore Digital Library and 33 (18.86% of total) of them were retrieved from Science Direct. This list of papers includes journal articles and conference proceedings.

The research group that executed this SLR is composed of one D.Sc. candidate and two professors, all addressing Artificial Intelligence and Data Mining topics. Fig. 3.1 shows the PRISMA flow diagram for this SLR. This diagram presents a systematic review's main activities, indicating the number of studies evaluated at each stage. The PRISMA flow diagram was proposed by Moher et al. [128] within a work that raised the preferred reporting items for systematic reviews and meta-analyses, the PRISMA statement.

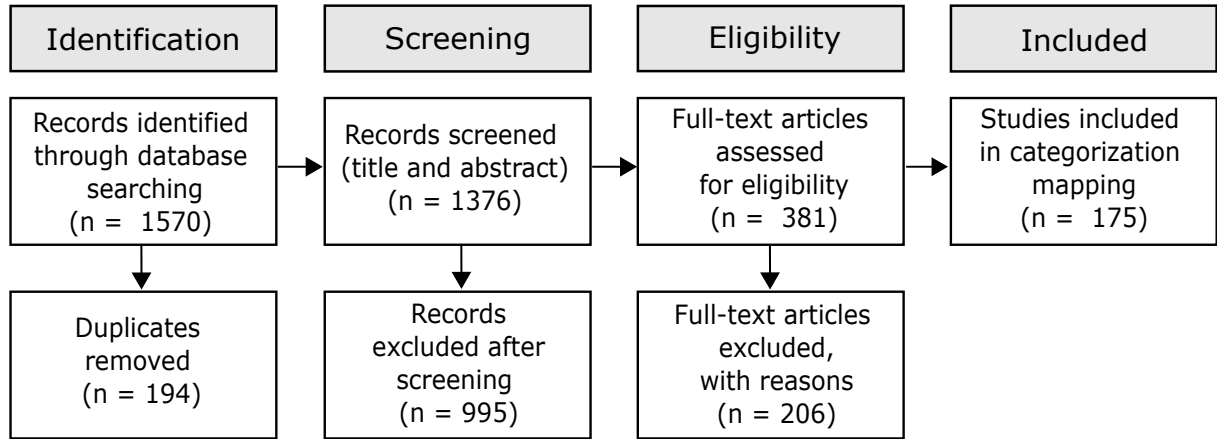


Figure 3.1: PRISMA flow diagram for this Systematic Literature Review (SLR).

## 3.2 Feature Selection Issues for Text Classification

We read and analyzed all included studies to identify the main issues that are being addressed by them (Research Question 1). After analyzing each study, we identified the

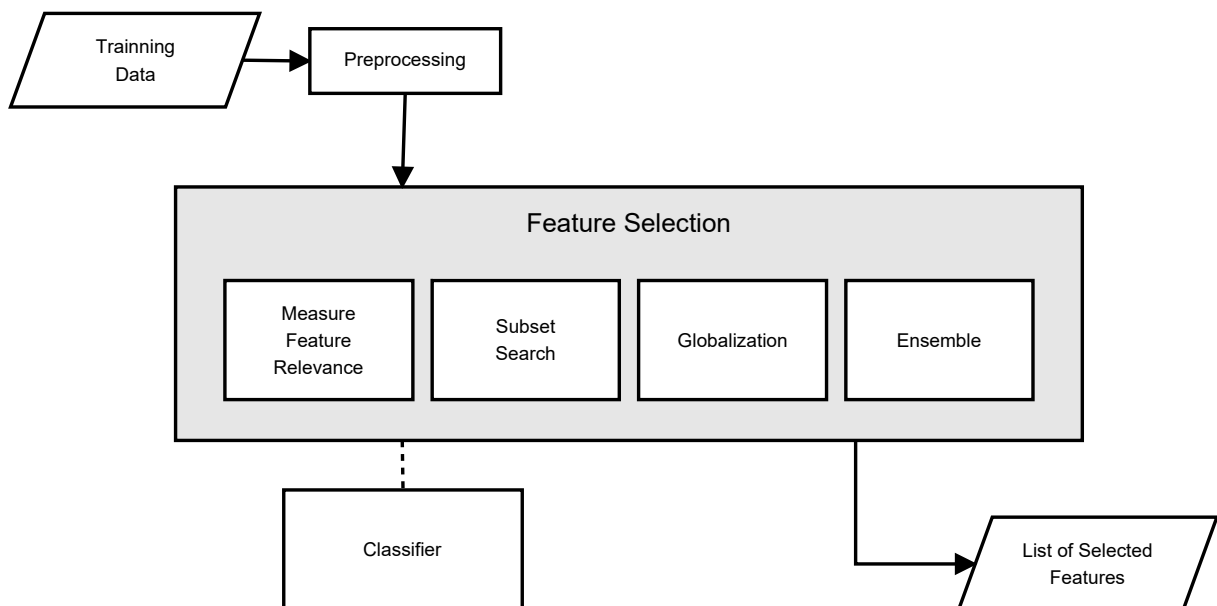


Figure 3.2: The four FS sub-tasks for text classification.

main groups of problems/issues and mapped the included studies to these groups. We found that these groups of problems represent sub-tasks of the FS process (Fig. 3.2). They are related to each other and can be organized as:

1. **Measure Feature Relevance** – Measure the relevance of each feature is an essential task in FS activity. There are different ways to estimate the relevance of features, such as measuring the correlation with the target, the variable entropy, or calculating the redundancy of features [85]. However, the basic idea is that the higher the relevance of a feature, the greater must be the power to increase the accuracy of the model (in our case, a text classifier). These studies compare existing metrics or define new metrics for calculating the predictive potential of each feature. The large part of the studies included in this review deal with issues related to the task of measure feature relevance.
2. **Subset Search** – The subset search task aims to find the best subgroup of features to be used in the classification. We found two main ways to perform this search: (a) evaluating several different subsets directly in the classification activity (wrapper method), and (b) using some heuristics to assess the relevance of each subset without evaluating in a specific classifier (filter method). In both approaches, optimization methods (such as genetic algorithms or Particle Swarm Optimization (PSO)) can be used to help the search. Subset search methods commonly use as its basis some of the existing feature relevance metrics (such as Chi-square (CHI), Information Gain (IG), or Mutual Information (MI)).
3. **Globalization** – Relevance metrics and subset search methods commonly can be applied specifically for one class or label of the dataset. Therefore, a method that globalizes the results of each class/label is required to construct a final set of features that represents all classes or labels. One alternative to globalization is to use specific sets of features for each class/label. However, the classifier must be designed to work this way. We mapped studies about class/label specific features in the globalization category in this review.
4. **Ensemble** – Each FS method has specific advantages and disadvantages, so combining two or more methods can lead to better results than using them separately. Ensemble studies propose or evaluate approaches to combining FS methods and/or metrics.

Sections 3.2.1 to 3.2.4 discuss the most relevant issues and studies for each task and present an overview of the main approaches we found to deal with each presented issue. The methods proposed in the included studies will be described in Section 3.3.

Fig. 3.3 presents the evolution of the number of studies by the issue group. This chart shows an increasing number of papers that address the subset search problem while a decreasing number of studies that focus on how to measure the relevance of features. Since the studies search for this review happened in October 2020, the numbers of papers for the 2020 are still partial because the source databases were still incomplete for those dates. Therefore, we mark this part of the graph in gray to show that it includes preliminary data.

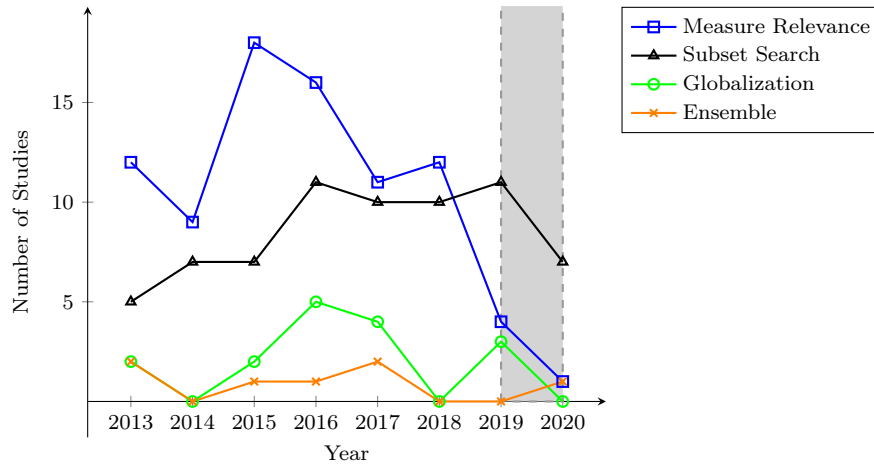


Figure 3.3: Number of studies by FS issue group over the years. The gray box indicates that data collected for 2020 may be preliminary since this review was updated in October 2020.

### 3.2.1 Issues About Measure Feature Relevance

After analyzing included papers that deal with the task of measure feature relevance, we mapped the most frequent issues related to this task:

- How to deal with the unbalanced or skewed datasets?
- How to avoid that rare terms receive high scores?
- How to identify and measure the redundancy of terms?
- How to consider the sparsity of the matrix?
- How to consider the position of terms?

Each one of these issues is explained next.

**Unbalanced or Skewed Datasets** For text classification, data are characterized by a large number of highly sparse terms and highly skewed categories [153]. A skewed dataset has an unbalanced class distribution, which means the number of instances in



one class (majority class(es)) may be many times bigger than the number of other class instances (minority class(es)) [70]. If the unbalance of classes is not treated during or before the FS, the positive features of the minority classes may receive minor relevance scores because they are present in a smaller number of documents. Consequently, the minority classes may have no features on the selected feature set. If some class/label does not have any of its features included in the final set, the classifier may not be able to classify the documents of this class/label.

**Rare Terms/Features** Several relevance metrics are based on the correlation between features and the target. In these metrics, features with high correlation receive higher scores. Because rare terms tend to be present in a few or only one class, they tend to have a higher correlation with the target. For this reason, many of the relevance metrics assign higher scores for rare features. As an example, we can cite Balanced Accuracy Measure (ACC2) and IG [152, 201]. However, rare features usually are not relevant to classification tasks, because they will have a very low likelihood of to be present in new documents.

We found two main strategies to deal with rare terms in included studies: (1) eliminating rare terms before applying the FS method; and (2) adapt the relevance metric formula to assign lower scores to rare terms.

**Redundancy Exclusion** Feature redundancy is usually defined as a high correlation among features [203]. Redundant features are likely to appear simultaneously in the same documents. Considering that two features are redundant and one of them is already selected, select the second feature will contribute very little or nothing to the total relevance of the selected set. For this reason, the redundancy between features is an important factor to measure the relevance of each feature. The goal of FS is to select a highly-relevant subset with a minimum redundancy [90].

Like the case of dealing with rare features, we found two main approaches that deal with redundant features: (1) eliminating redundant features during the pre-processing phase before applying the FS method; and (2) adapt the relevance metric formula to assign lower scores to redundant features.

**Data Sparsity** For text classification, features usually consider in its recipe the frequency of words/phrases in each document. The number of possible words/phrases in

all dataset training documents can be huge, and this number defines the length of the initial feature vector [135]. Since each document usually includes a low percentage of the total set of words/phrases, most of the features of each document will be zero, i.e., zero frequency for words/phrases that are not present in that document. The result is a sparse composite matrix.

The matrix sparsity degrades the performance of the text classification [135]. For this reason, properly handling the sparsity issue is recommended for FS methods. An alternative to address this issue is to use another representation format which does not result in a sparse matrix.

**Position/Location Inside the Text** The term location inside a text can be related to the importance of a term and, consequently, for measuring the relevance of a feature. Song et al. [174] defines that the feature words have different capabilities to express the text in different positions of the text. Especially for news articles, information that is in the title, subtitle, or first paragraph tends to be more relevant than information on the other paragraphs. Therefore, the location of terms within the text can be useful for FS methods.

### 3.2.2 Issues About Subset Search

The subset search task aims to find the best subgroup of features to be used in the classification. This search can be performed using an optimization method (such as genetic algorithms or PSO) and evaluating several different subsets directly in the classification activity (wrapper method) or using some heuristics to evaluate the relevance of each subset without evaluating in a specific classifier (filter method). Most subset search studies are focused on evaluating metaheuristics methods to improve search efficiency. Other studies focus on reducing redundancy and hyperparameter optimization.

**Search Strategy/Search Efficiency and Effectiveness** The simplest way to perform the subset search is exhaustively to evaluate all candidate subsets according to some evaluation function. However, for a dataset with  $N$  features, there exists  $2^N$  candidate subsets. Due to a large number of features in text domain datasets, the exhaustive search usually is too costly and virtually prohibitive [187]. For this reason, the main issue of the subset search for text classification is the search efficiency. Several studies included in this review proposes different metaheuristics or methods to perform the search efficiently.

Most of the included studies about subset searches are based on swarm optimization methods. In those methods, subsets of candidate features are mapped as particles, and the Swarm Optimization method tries to find the best solution (best feature subset) by exploring the search space moving the particles to find the global optimum configuration [37].

**Redundancy** Feature redundancy is usually defined in terms of some correlations within the features [203]. The goal of FS is to select a highly-relevant subset with a minimum redundancy [90]. Most subset search methods avoid select redundant features naturally by evaluating several combinations of features. As explained in Section 3.2.1, methods that measure the relevance of each feature in isolation need address feature redundancy explicitly.

**Feature Selection and Hyperparameter Optimization** Classification algorithms usually have several parameters whose values heavily influence its performance. Thus, determining appropriate values of parameters of a classifier is a critical issue [41]. Like FS, finding the best combination of parameters can be addressed as an optimization problem, called Hyperparameter Optimization. Grid search and manual search are the most widely used strategies for hyperparameter optimization [25].

Subset search methods usually are wrapper methods. That is, they use the accuracy of the classifier to evaluate each candidate subset. But notice that the predictor parameters influence the performance of the subset search. Similarly, the selected features influence hyperparameter optimization. For this reason, performing the two searches in an integrated manner may be a good approach to find the best combination of selected features and predictor parameters.

### 3.2.3 Issues About Globalization

Relevance metrics and subset search methods commonly can be explicitly applied for one class or label of the dataset. Therefore, a method that globalizes the results of each class/label is required in order to construct a final set of features that represents all classes or labels. One alternative to globalization is to use specific sets of features for each class/label. Studies about class/label specific features are mapped in the globalization category in this review.

Analyzing included studies, we found three ways to implement FS globalization within

a text classification architecture:

1. Implement a local FS method for each class/label and perform the globalization subsequently.
2. Implement a global FS method designed to deal with globalization problems.
3. Adapt/Use a class/label specific classification scheme (selecting specific features for each class/label).

For the first three approaches, the main issue to be addressed is the representativeness of each class and label in the selected final set of features. For the fourth approach, the main question is how to transform the classifier or transform the problem to be able to use specific subsets of features per class/label. These globalization issues will be detailed in the following two paragraphs.

**Classes/Labels Representativeness on Final Feature Set** The basic scheme of filter-based FS assigns a score to each feature based on its discriminating power. It selects the top- $k$  features from the feature set, where  $k$  is an empirically determined number [4]. If the classification problem is multiclass or multi-label, some classes/labels may have few or no selected features in the final feature set. Therefore, a central issue for globalization is ensuring adequate representativity for all classes/labels in the final dataset.

**Class/Label Specific Features** Instead of performing the globalization and obtaining a single subset of features to be used in the classifier, it is possible to use subsets of specific features for each class or label. Some studies show this approach can improve the classification performance [186, 184]. However, the classical theory as it stands requires operating in a common feature space and fails to provide any guidance for a suitable class-specific architecture [15]. Therefore, when using class/label specific features, the central issue is how to adapt the problem or the classifier to work with these class/label specific features.

In addition to the globalization approaches we identified in our systematic review, other approaches that address the globalization issue:

1. FS based on sparse learning, which focuses on the relationship between features and classes or labels [27].
2. FS based on manifold learning [216].

### 3.2.4 Issues About Ensemble

Each FS method has specific advantages and disadvantages, so combining two or more methods can lead to better results than using them separately. Ensemble studies propose or evaluate approaches by combining FS methods or metrics. We found that only seven of the included studies address the issue of ensembling FS methods. Another included studies deal with the FS methods specifically for ensembling learning approaches (for example, Boosting-based algorithms [7]), but they are not focused on ensembling FS methods. Virtually all of the included studies about the ensemble issue address the same central problem of how to combine and aggregate the results of different FS methods. We found three main approaches to ensemble FS methods:

1. Combining selected subsets – Execute/Performs two or more FS methods isolated and then create a final set of features by combining the subsets selected by the different methods.
2. Chaining FS methods – Execute two or more FS methods in sequence, where the subset selected by a method becomes the input for the next method.
3. Ensembling rankings – Construct two or more relevance rankings (using different relevance metrics), combining the resulting rankings into a unique ranking and finally select the features using a predetermined threshold.

For each approach, different ways of combining or rankings the sets are available. The main issue of the included studies is precisely to define/find the best way to perform this combination to obtain the most relevant subset of features.

## 3.3 Feature Selection Methods for Text Classification

As explained in Section 3.2, FS for text classification should address different issues. Our SLR found that several types of methods are being proposed and evaluated to address such issues. We analyzed all the included studies of our SLR and mapped the main characteristics of each method. Based on this mapping, we designed a new categorization scheme that allows to group the methods from four different perspectives (Research Question 2): strategy, approach, target, and labeled data dependence (Fig. 3.4). The proposed categorization scheme helps to organize in groups and compare current methods. Additionally, it will help the positioning of future studies about FS for text classification.

The first perspective addresses the different strategies as the selection of features

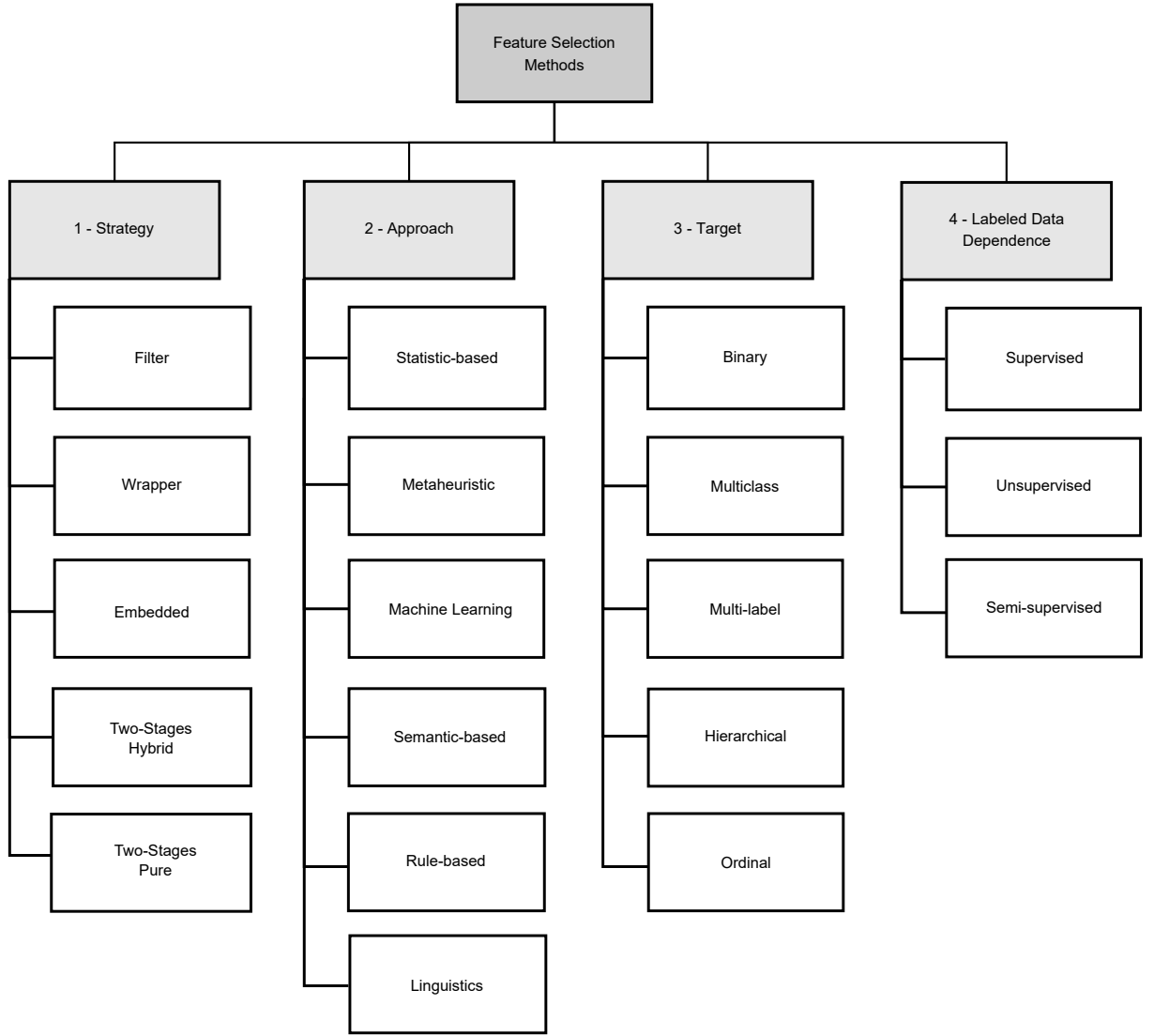


Figure 3.4: Proposed categorization schema for FS methods for text classification. Each vertical represents a different categorization perspective. We used this categorization scheme to map and analyze the 175 FS methods included in this review.

can be performed. It is detailed in Section 3.3.1. The different approaches (statistical, machine learning, or semantical) are mapped on the second perspective and is detailed in Section 3.3.2. The third perspective maps the type of target that the method was built to handle (binary, multiclass, multi-label, hierarchical, or ordinal). It is explained in Section 3.3.3. The fourth perspective maps the level of dependence on labeled data, being detailed in Section 3.3.4.

Each perspective is composed of a set of categories, as shown in Fig. 3.4. We mapped each of the studies included in this review according to each of these four perspectives of the classification schema applied on the methods described by them. Table 3.1 maps the issues groups described in Section 3.2 into each perspective of our categorization

schema presented in this section. This table can be used to identify which strategies and approaches are being used to address each issue group. The most relevant studies on each category will be indicated during the explanation of the perspectives and respective categories in the following sections.

Table 3.1: Number of FS studies for text classification by issue group and categorized according to the proposed categorization schema. All included studies were published between January/2013 and October/2020. Note: We use some abbreviations to simplify this table. TS refers to Two-Stages, and ML refers to Machine Learning.

Issue Group	Strategy	Approach	Target	Labeled Data Dependence
Measure Relevance (84 Studies)	Filter (77)	Statistic-based (63)	Multiclass (59)	Supervised (77)
	TS Pure (6)	Semantic-based (9)	Binary (19)	Semi-superv. (4)
	Embedded (1)	ML (7)	Hierarchical (3)	Unsupervised (3)
		Linguistics (3)	Ordinal (2)	
Subset Search (68 Studies)	Wrapper (22)	Rule-based (2)	Multi-label (1)	
	Filter (21)	Metaheuristic (30)	Multiclass (43)	Supervised (66)
	TS Hybrid (16)	Statistic-based (25)	Binary (19)	Semi-superv. (1)
	Embedded (5)	ML (10)	Multi-label (6)	Unsupervised (1)
Globalization (16 Studies)	TS Pure (4)	Semantic-based (2)		
	Filter (14)	Rule-based (1)		
	Embedded (1)	Statistic-based (16)	Multiclass (14)	Supervised (16)
Ensemble (7 Studies)	Wrapper (1)		Multi-label (2)	
	Filter (5)	Statistic-based (5)	Binary (5)	Supervised (7)
	TS Hybrid (1)	Metaheuristic (1)	Multiclass (2)	
<b>Total (175 Studies)</b>	Wrapper (1)	Rule-based (1)		
	Filter (117)	Statistic-based (109)	Multiclass (118)	Supervised (166)
	Wrapper (24)	Metaheuristic (31)	Binary (43)	Semi-superv. (5)
	TS Hybrid (17)	ML (17)	Multi-label (9)	Unsupervised (4)
	TS Pure (10)	Semantic-based (11)	Hierarchical (3)	
	Embedded (7)	Rule-based (4)	Ordinal (2)	
		Linguistics (3)		

### 3.3.1 Categorization by Strategy

As presented in Chapter 2, FS methods are usually classified into three categories: filter, wrapper, and embedded. The first three flows in Fig. 3.5 represents each one of these strategies. In this SLR, we have included two more categories to the three classical ones: Two-stages Pure and Two-stages Hybrid. Note that both strategies have the same flow in Fig. 3.5. The difference is in the choice to combine the same or different strategies. Two-stages Hybrid strategy methods combine FS methods are based on different strategies of selection. For example, the first stage may apply the filter strategy, while the second stage may use the wrapper strategy. On the other hand, some studies combine two different methods but using the same strategy. For these cases, we classify the studies in a separate category (Two-stages Pure strategy). Each one of the five strategies considered in this survey will be presented next. Fig. 3.6 summarizes the number of included studies by strategy over the years.

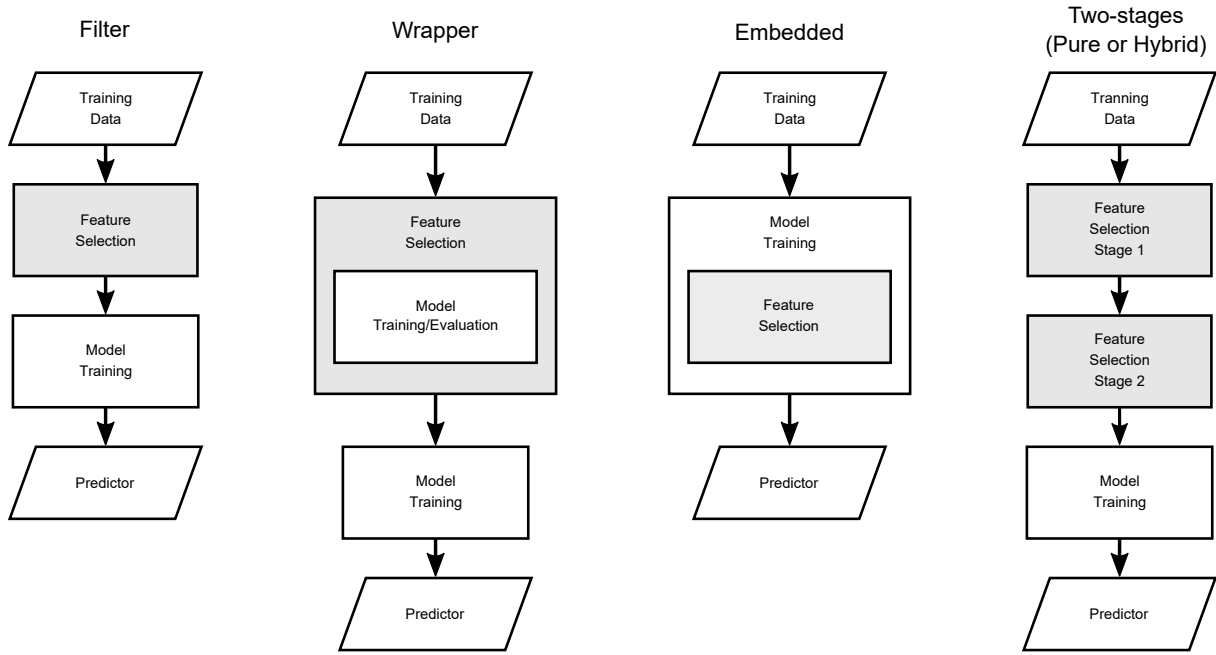


Figure 3.5: Flow diagram for each Feature Selection (FS) Strategy presenting the interaction between the FS activity and the model training activity. Each of these strategies is detailed in Section 3.3.1.

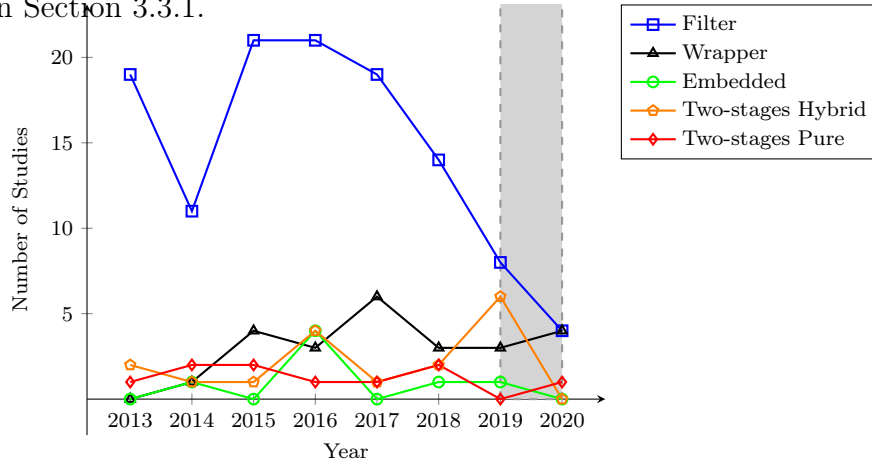


Figure 3.6: Number of FS studies by strategy over the years. The gray box indicates that data collected for 2020 may be preliminary since this review was updated in October 2020.

**Filter Strategy** The main characteristic of the methods that are based on the filter strategy is to be independent of the classifier. In other words, the filter strategy does not use the performance of the classifier to assess the relevance of features or subsets of features. Lazar et al. [96] subdivided these filter methods into two classes: ranking-based and space search. Ranking-based methods use some relevance metric to assess the predictive power of each feature, construct a ranking based on this relevance score, and apply a threshold to select the most relevant features [30]. Space search methods aim to find the best subset of features by evaluating different combinations of features.

Table 3.2 summarizes studies that apply a ranking-based approach to implement



Table 3.2: Filter methods for measure relevance.

Base Method	Studies
Accuracy Measure (ACC)	[152, 153]
Chi-square (CHI)	[50, 2, 18, 180, 17]
Class Discriminating Measure (CDM)	[52]
Cluster-based	[217, 172, 131, 118, 38, 58]
Comprehensively Measure Feature Selection (CMFS)	[43, 230]
Crowd-based Feature Selection (CrowdFS)	[144]
Discriminative Features Selection (DFS)	[235]
Document Frequency (DF)	[36, 110, 206, 104, 102, 103, 166, 229, 231]
Entropy	[127, 210]
Gini Index (GI)	[36, 212, 136]
Information Gain (IG)	[35, 140, 225, 51, 72, 135, 211, 233, 150]
Latent Dirichlet Allocation (LDA)	[8, 234]
Matrix	[112, 200]
Mutual Information (MI)	[16, 35, 73, 213, 55, 113]
Ontology-based	[146]
Part of Speech Filter (POSFilter)	[72, 147]
Position-based	[174]
Relative Discrimination Criterion (RDC)	[154, 90]
Rule-based	[172, 2, 137]
Student's <i>t</i> -Test	[145, 201]
Term Frequency-Inverse Document Frequency (TF-IDF)	[107, 59]
Word Embedding	[160, 233, 189, 93]
Other Methods	[11, 61, 108, 109, 155, 203, 12, 13, 206, 219, 63, 138, 158, 190, 193, 204, 207, 121, 122, 69, 79, 202]

the filter strategy. The studies in this table are grouped according to the base method employed to handle the problem of measuring the relevance. We found that these studies usually propose an improved version of some existing relevance metrics or propose new relevance metrics. As detailed in Section 3.1, our systematic review focused on mapping studies published between 2013 and 2020. However, some studies published before this time window have proven themselves in the literature. For example, the Distinguishing Feature Selector (DFS\*) method proposed by Uysal and Gunal [195].

However, filter methods are not restricted to handle the problem of measure relevance (ranking methods). The filter strategy can also be used to handle the subset search, globalization, or ensemble problems. Table 3.3 presents the filter methods that address each of these issues. In Table 3.3, the studies are also grouped according to the base method utilized.

As described in Section 3.2.1, one of the most relevant issues for text classification is data sparsity. The FS methods generally help to reduce data sparsity by removing less relevant features. However, we found one study based on the filter strategy that mainly

Table 3.3: Filter methods for subset search, globalization, and ensemble issues.

Problem	Base Method	Studies
Subset Search	Clustering	[175, 232, 159]
	Firefly Algorithm	[95]
	Geometric Properties	[177]
	Harmony Search	[205]
	Maximum Discrimination	[185]
	Mutual Information	[188, 67]
	Particle Swarm Optimization (PSO)	[76, 222]
	Rough Set	[88, 236, 29]
	Support Vector Machines (SVM)	[162]
	Syntax Features	[196]
	Word Embedding	[179, 221, 92]
	Other Methods	[106, 111, 191, 60]
Globalization	At Least One FeaTure (ALOFT)	[143, 48]
	Chi-square (CHI)	[214]
	Class Specific Features	[186]
	Global Filter-based Feature Selection Scheme (GFSS)	[194, 4, 5]
	Information Gain (IG)	[169, 215, 66]
	Mutual Information (MI)	[97, 3]
	Other Methods	[6, 218]
Ensemble	Blended Feature Selection Method (BFSM)	[171]
	Genetic Rank Aggregation	[134]
	Hybridized term-weighting	[163]
	Meta Feature Selection (MFS)	[105]
	Square of Information Gain and Chi-square (SIGCHI)	[62]

focuses on dealing with this issue. Ong et al. [135] propose an improved FS metric known as Sparsity Adjusted Information Gain (SAIG), which modifies the conventional IG metric and aims to adjust the feature ranking scores according to the matrix sparsity.

Additionally, some two-stage FS methods perform the filter strategy in both stages (Two-stages Pure Strategy). Despite having two stages, these methods cannot be classified as having a hybrid strategy because they only use one strategy (the filter strategy). As an example, we can cite the study of Karabulut [76] that presents a novel two-stage filter method based on IG theory and Geometric Particle Swarm Optimization (GPSO).

**Wrapper Strategy** Different from the filter strategy, the methods that use the wrapper strategy are dependent on the predictor because they use the performance of the predictor to evaluate the relevance of features or search for the best subset of features. For this reason, wrapper methods tend to be more computationally costly than filter methods. Most wrapper methods are based on search techniques. As explained in Section 3.2.2, the exhaustive search usually is too costly and most times prohibitive [187]. For this reason, the main issue related to wrapper methods is the search efficiency (ex-

plained in Section 3.2.2). Several studies included in this review propose different methods to improve search efficiency using metaheuristic search methods. Therefore, we identified that the wrapper strategy is commonly implemented with a metaheuristic approach. Section 3.3.2 details the metaheuristic approach and lists the included studies based on this approach. Wrapper methods usually are subset search methods. That is, they use the accuracy of the classifier to evaluate each candidate subset. Therefore, the predictor parameters influence the subset search performance. Similarly, the features selected influence hyperparameter optimization. For this reason, wrapper methods perform the two searches (subset search and hyperparameter search) in an integrated manner can be the best approach to find the best combination of selected features and predictor parameters. Despite the relevance of this issue, the only study included addressing it was developed by [41].

**Embedded Strategy** The main characteristic of embedded methods is the incorporation of the FS as part of the training process. Embedded methods aim to reduce the computation time taken up for reclassifying different subsets, which is done in wrapper methods [30]. The embedded methods include studies that are evaluated in specific/atypical learning situations:

- Aspect-based Sentiment Analysis – [224].
- Class-specific Features – [186].
- Ensemble of Multi-label Classifiers – [56].
- Multi-objective Genetic-Programming – [129].
- Positive and Unlabeled Learning – [228].

Our review found that most embedded strategy studies (5 of 7) focus on the subset search issue described in Section 3.2. As embedded FS methods are part of the training algorithm, this strategy can deal efficiently with the subset search issue [129]. Additionally, we identified one embedded strategy study focused on the measure relevance issue [130] and another one focused on the globalization issue by implementing class-specific features [186]. None of the embedded strategy studies in this review focus on FS ensemble issue.

**Two-stages Hybrid Strategy** Each of the strategies presented until now (filter, wrapper, and embedded) has specific advantages and disadvantages. For this reason,

many studies explore hybrid methods that combine two different strategies into a single method. In this way, it is possible to combine their advantages and mitigate specific problems/risks. Several studies perform a filter stage before conduct the subset search to reduce the search space. For example:

- Filter Stage + Genetic Algorithm Based Search – [53].
- Filter Stage (IG or CHI) + Rough Set – [88].
- Filter Stage + Markov Blanket Filter (MBF) Subset Search – [71].
- Filter Stage + Support Vector Machine-Recursive Feature Elimination (SVM-RFE) – [226].
- Filter Stage (CHI) + Spark BAT Feature Selection (SBATFS) – [32].
- Filter Stage (CHI) + Particle Swarm Optimization (PSO) – [173].
- Filter Stage (MI) + Recursive Feature Elimination (RFE) – [74].
- Filter Stage (IG) + Particle Swarm Optimization (PSO) – [19].
- Filter Stage (IG) + Binary Gravitational Search Algorithm (BGSA) – [77].
- Filter Stage (IG) + Improved Sine Cosine Algorithm (ISCA) – [22].
- Filter Stage (Ontology Filter) + Particle Swarm Optimization (PSO) – [1].

**Two-stages Pure Strategy** Some studies combine two different methods but using the same strategy. For this reason, they cannot be classified as hybrid strategy methods. Therefore, we classify them into a different strategy (Two-stages Pure). The following studies combine two stages based on filter strategy or based on wrapper strategy:

- Filter (IG or CHI or MI) + Filter (Clustering) – [53].
- [109] propose a two-step FS method. At the first step, redundancy analysis among original features based on a categorical fuzzy correlation degree is applied to filter the redundant features with a similar categorical term frequency distribution. In the second step, a conventional IG feature relevance metric is adopted to select the final feature set.
- Wrapper (Forward Feature Construction) + Wrapper (Genetic Algorithm) – [149].

### 3.3.2 Categorization by Approach

During the review, we identified that FS works could be grouped according to the approach used. In this SLR, the approach is related to the computational, statistical, or

semantic technique used to select features. While the strategy (Section 3.3.1) defines how the method will fit into the training process and how it relates to the classifier, the approach (presented in this section) concerns the technique employed to perform the selection of features. We decided to map each method based on their primary approach since we identified that most methods do some combination of approaches, mainly with the statistic-based approach. For this reason, we did not map them into a separate category (hybrid approach).

Most published studies (109 of 175, 62.29%) use statistical metrics to measure the relevance of features and select them. These methods will be classified as statistic-based approaches. However, other studies use different approaches to select features. The main groups of approaches we have found were machine-learning-based techniques (such as clustering), semantic-based techniques, and rule-based techniques (such as Apriori). Each of these approaches will be detailed below.

**Statistic-Based Approaches** Gunduz and Cataltepe [55] propose a feature relevance metric called Balanced Mutual Information (BMI) that is able to deal with the class imbalance problem through oversampling of the minority classes. They use the Synthetic Minority Oversampling Technique (SMOTE) for oversampling, which creates new minority class instances by searching for nearest neighbors of a randomly selected minority class instance. The new minority class instance value is generated by interpolation of randomly selected instances and selected neighbors of this instance. Rehman et al. [153] propose a new feature relevance metric called Max-Min Ratio (MMR). It is a product of max-min ratios of true positives and false positives and their difference, which allows MMR to select smaller subsets of more relevant terms even in the presence of highly skewed classes.

As discussed in Section 3.2.1, there are two main strategies to deal with rare terms: (1) eliminate rare terms during the pre-processing phase before applying the FS method; and (2) adapt the relevance metric formula to assign lower scores to rare terms. Rehman et al. [154] adopt the first strategy explicitly by removing rare features before evaluating the proposed relevance metric called Relative Discrimination Criterion (RDC). Rehman et al. [152] adopt the second strategy in a recent study. They propose the Normalized Difference Measure (NDM) that is an improved version of the ACC2 [47] modified specially to assign lower relevance scores to rare features.

Labani et al. [90] demonstrated that RDC is an effective method for identifying relevant features. A drawback is that the correlation between features is ignored, and thus

RDC cannot identify redundant features. In order to mitigate this problem, Labani et al. [90] propose the Multivariate Relative Discrimination Criterion (MRDC) that is an evolution of the RDC. Labani et al. [90] modified the original formula to identify and measure the redundancy of features based on the correlation between them. As a result, MRDC assigns a higher relevance score to features with high discriminative power and low redundancy.

Document Frequency (DF) of a feature refers to the number of documents that include that feature. The term frequency refers to the occurrence number of a certain feature in a certain document. Most popular FS metrics for text classification such as IG, CHI, and Odds Ratio (OR), are based on DF and don't use the term frequency [11]. However, the term frequency is a piece of important information for FS because it represents the importance of feature to each document [211]. High-Frequency terms (except stop words) that occurred in few documents are often regarded as discriminators in the real-life corpus [201].

To overcome this drawback, Baccianella et al. [11] propose to logically break down each training document of length  $k$  into  $k$  training "micro-documents", each consisting of a single word occurrence and endowed with the same class information of the original training document. This transformation has the double effect of (a) allowing all the original FS methods based on binary information to be still straightforwardly applicable, and (b) making them sensitive to term frequency information. Wang et al. [201] propose a new FS metric based on term frequency and Student's  $t$ -Test. The  $t$ -Test function is used to measure the diversity of the distributions of a term frequency between the specific category and the entire corpus. Wu and Xu [211] propose a new FS metric that combines DF and term frequency called Limiting DF's Word Frequency. Its primary principle is summarized as follows: pre-set the threshold value of minimum DF  $\alpha$  and the threshold value of maximum DF  $\beta$ , if the DF of feature word is between  $\alpha$  and  $\beta$  then calculate the word frequency of this feature word or delete it otherwise.

**Metaheuristic Approaches** As explained in Section 3.3.1, metaheuristic search methods can be implemented to address the subset search issue and usually is combined with wrapper strategy. Metaheuristic algorithms use problem-specific heuristic information and efficiently manage the search process without exploring the whole search space [54]. Therefore, they are ideal candidates to overcome the drawbacks of wrapper-based methods [54]. Common meta-heuristic algorithms include the genetic algorithm

and PSO [114]. The included studies that implement the metaheuristic approach is listed below:

- Binary Black Hole Algorithm (BBHA) – [139].
- Binary Particle Swarm Optimization (BPSO) – [170].
- Cat Swarm Optimization (CSO) – [114].
- Genetic Algorithm and Wrapper Approaches (GAWA) – [149].
- Improved Particle Swarm Optimization (IPSO) – [117].
- Multi-Objective Automated Negotiation based Online Feature Selection (MOANOFs) – [24].
- Multi-Objective Relative Discriminative Criterion (MORDC) – [91].
- Memetic Feature Selection based on Label Frequency Difference (MFSLFD) – [99].
- Optimized Swarm Search-based Feature Selection (OS-FS) – [46].
- Small World Algorithm (SWA) – [116].
- Wrapper Feature Selection Algorithm based on Iterated Greedy (WFSaIG) – [89].
- Wolf Intelligence Based Optimization of Multi-Dimensional Feature Selection Approach (WI-OMFS) – [54].

**Machine Learning-Based Approaches** Among the included studies, 17 studies use some machine learning methods directly in the FS process. These studies mainly used the following techniques:

- Clustering – [175, 217, 232, 172, 131, 159, 118, 38, 84, 58].
- SVM – [162, 226, 191].
- Word Embedding – [221, 92, 189, 93].

**Semantic-Based Approaches** Evaluate the meaning of words can be useful for FS methods because it helps to identify the relevance of words inside a text and identify the similarity between words. Among the studies included, only 11 studies use a semantic approach. Below are the semantic technologies used by each study:

- Context-capturing Features – [61];
- Crowd-based Feature Selection (CrowdFS) – [144].
- Discriminative Personal Purity (DPP) – [136].
- Latent Selection Augmented Naive Bayes (LSAN) – [42].
- Ontology – [146, 1].
- Semantic Measures – [137].

- Semantic Similarity – [235].
- Word Embedding – [179, 233].
- Topic Guessing – [122].

We categorize two studies [179, 233] that use Word Embeddings as a semantic approach because it was used to map the meaning of the words. Both studies used Word Embedding to map the similarity of the words and perform the similarity expansion in FS. The aim of similarity expansion is to expand the set of selected features based on similarity of words.

**Rule-Based Approaches** Among the studies included, only four use rule-based approach. Agnihotri et al. [2] propose a novel hybrid FS called Correlative Association Score (CAS) of terms. The CAS utilizes the concept of the Apriori algorithm to select the most informative terms. Sheydaei et al. [172] proposed the Bit-priori Association Classification Algorithm (BACA), which combines the rule approach with a semantic approach. More recently, Wang and Hong [202] proposes the Hebb Rule Based Feature Selection (HRFS) that assumes that terms and classes are neurons and select terms under the assumption that a term is discriminative if it keeps “exciting” the corresponding classes. Finally, Sundararajan et al. [181] proposes the multi-rule based ensemble FS model for sarcasm classification.

**Linguistics Approaches** In our review, we found three FS studies based mainly on the linguists’ approach. The proposed methods use lexical or grammar information to measure the relevance of the features [127, 147, 72].

### 3.3.3 Categorization by Target

Document classifiers may have different types of targets. Binary classifiers estimate one class for each new document within two possible categories (usually positive and negative categories). Multiclass classifiers assign each new document to one class from a list including three or more possible classes. In multi-label classification, a classifier attempts to assign multiple labels to each document, whereas a hierarchical classifier maps text onto a defined hierarchy of output categories [126]. Hierarchical and ordinal classifiers can be viewed as specific types of multiclass classifiers in which classes have a relationship with



Table 3.4: FS studies using binary target text datasets.

Problem Domain	Studies
Sentiment Analysis	[16, 35, 155, 88, 179, 112, 135, 127, 138, 170, 134, 168, 191, 223, 84, 173, 54, 69]
Spam Detection	[10, 206, 115, 148, 122]
Other Domains	[61, 171, 203, 13, 226, 228, 114, 137, 163, 166, 118, 152, 234, 121, 153, 1, 198]

each other. In the hierarchical classification, the classes are organized into hierarchical levels, whereas in ordinal classification, the classes are organized in order or sequence.

During our review, we identified that each FS method is specifically designed to work with a specific target type. The following paragraphs present the main proposed methods of each target type.

**Binary Text Classification** Among the 175 included studies, 43 studies (24.57%) focus on FS for binary text classification. We found that 20 studies (46.51% of 43) are related to the sentiment analysis and 6 studies (13.95% of 43) are related to spam detection. Both sentiment analysis and spam detection are usually handled as a binary classification problem. Table 3.4 presents the FS studies that were evaluated with binary datasets grouped by problem domain.

**Multiclass Text Classification** Most of studies about FS for text classification (118 of 175 included studies, 67.43%) focus on multiclass. Among these, 82 studies (69.49% of 118 studies) evaluate the method proposed using the main news classification benchmarks (datasets Reuters-21578, 20Newsgroup, Fudan, Sogou News). FS methods for multiclass or multi-label text classification need to address the globalization issue described in Section 3.2.3. In our review, we found studies for each implementation options described in Section 3.2.3:

1. Implement a local FS method for each class/label and subsequently perform globalization [169, 215].
2. Implement a global FS method designed to deal with globalization problems [97, 4].
3. Adapt/Use a class/label specific classification scheme [184, 186].

**Multi-Label Text Classification** Among the included studies, only nine studies focused on FS for multi-label text classification:

- Based on supervised topic modeling for Boosting-based multi-label text categorization – [8].
- Using Diversified Greedy Backward-Forward Search (DGBFS) – [161].
- Using Ensemble Embedded Feature Selection (EEFS) – [56] and [57].
- Using label Pairwise Comparison Transformation (PCT) method, which converts each original multi-label sample into multiple samples with same feature vectors and different label vectors – [214].
- Using Multivariate Mutual Information (MMI) – [97].
- Using two-stage term reduction strategy based on IG theory and GPSO search – [76].
- Using Fuzzy Rough Feature Selection (FRFS) – [236].
- Using Memetic Feature Selection based on Label Frequency Difference (MFSLFD) – [99].

As detailed in Section 3.1, our SLR protocol focused studies that explicitly state the application of text classification in the title or abstract. However, some works outside this scope are also interesting and were experimentally tested on text data. A relevant example is the mutual Information-based multi-label FS method using interaction information [98].

**Hierarchical Text Classification** Among the included studies, only three of them focused on FS for hierarchical text classification. Naik and Rangwala [130] investigate various filter-based FS methods for dimensionality reduction to solve the large-scale hierarchical classification problem. Lifang et al. [113] propose a hierarchical FS method using Kullback-Leibler divergence to measure the correlation between the class and subclasses, and using MI to calculate the correlation between each feature and subclass. Song et al. [174] propose a FS method based on category distinction and feature position information for Chinese text classification. This is the only included study in our review that deals with the issue of considering the position of words during the FS process.

**Ordinal Text Classification** Among the included studies, only two focused on FS for ordinal text classification. Baccianella et al. [11] evaluate the use of micro-documents in ordinal classification. They logically break down each training document of length  $k$  into  $k$  training “micro-documents”. The purpose of the use of micro-documents was explained earlier in this section. Baccianella et al. [12] propose four novel FS metrics that have been specifically devised for ordinal classification and test them on two datasets of product review data.

### 3.3.4 Categorization by Labeled Data Dependence

According to the Encyclopedia of Machine Learning [165], supervised learning refers to any machine learning process that learns a function from an input type to an output type using data comprising examples that have both input and output values. The same Encyclopedia, define unsupervised learning to any machine learning process that seeks to learn structure in the absence of either an identified output and semi-supervised learning to any machine learning process that uses both labeled and unlabeled data to perform an otherwise supervised learning or unsupervised learning task. Labeled data are data for which each object has an identified target value, the label [165].

Like learning methods (such as classification and regression), FS methods can also be classified into supervised, unsupervised, and semi-supervised according to their dependence on labeled data. FS methods that need labeled data can be classified as supervised method. On the other hand, FS methods that don't need labeled data can be classified as an unsupervised FS methods. Finally, FS methods that work with both labeled and unlabeled data are classified as semi-supervised.

**Supervised Methods** Most FS studies for text classification propose supervised methods. Considering the 175 studies included in this review, 166 (94.86% of total) are based on supervised methods. Supervised methods are mostly methods that measure the relevance of features alone or in subsets of features based on a labeled training set. Table 3.5 present all included studies grouped by labeled data dependence and year of publication.

**Unsupervised Methods** Considering the 175 studies included in this review, only four (2.29% of total) are based on unsupervised methods. In these works, three different unsupervised techniques were used:

- Term Frequency-Inverse Document Frequency (TF-IDF) and Glasgow expressions – [121] propose two modifications to the traditional TF-IDF and Glasgow expressions using graphical representations to reduce the size of the feature set.
- Word Co-occurrence Matrix – [200] propose an unsupervised FS algorithm through Random Projection and Gram-Schmidt Orthogonalization (RP-GSO) from the word co-occurrence matrix.
- Word Embedding – [160] propose an unsupervised FS method that utilizes Word Embedding to find groups of words with similar semantic meaning. Word Embed-

Table 3.5: FS studies grouped by labeled data dependence and year of publication.

Type	Year	Studies
Supervised	2013	[10, 11, 16, 35, 61, 68, 73, 76, 97, 105, 108, 109, 140, 162, 169, 171, 175, 177, 213, 225, 203]
	2014	[13, 12, 36, 51, 88, 110, 145, 161, 179, 206, 205, 217, 222, 226, 232]
	2015	[41, 42, 43, 50, 55, 62, 71, 72, 107, 112, 117, 135, 143, 154, 172, 215, 219, 235]
	2016	[2, 18, 45, 46, 48, 53, 102, 103, 106, 111, 114, 115, 127, 129, 130, 131, 137, 138, 147, 158, 159, 163, 166, 170, 174, 185, 186, 183, 190, 194, 210, 211, 221, 229, 230]
	2017	[3, 4, 8, 34, 49, 56, 66, 87, 113, 116, 118, 134, 144, 148, 152, 168, 180, 191, 196, 204, 212, 214, 223, 233, 234]
	2018	[19, 23, 28, 33, 38, 52, 60, 59, 84, 90, 95, 133, 136, 146, 150, 153, 17, 189, 224, 231, 236]
	2019	[1, 5, 6, 22, 32, 57, 69, 74, 77, 79, 89, 99, 122, 173, 188, 198, 202, 218]
	2020	[54, 91, 93, 24, 29, 58, 67, 149, 181]
Unsupervised	2016	[160, 200]
	2017	[92]
	2018	[121]
Semi-Supervised	2013	[155]
	2014	[228]
	2016	[63, 193]
	2017	[207]

ding maps the words into vectors and remains the semantic relationships between words. After mapping the similar semantic groups, the method maintains the most representative word on behalf of the words with similar semantic meaning. [92] propose an unsupervised FS method that uses Neural Word Embeddings, trained on social media content from Twitter, to determine how strongly textual features are semantically linked to an underlying health concept.

**Semi-Supervised Methods** Semi-supervised learning uses both labeled and unlabeled data to perform an otherwise supervised learning or unsupervised learning task [165]. Considering the 175 studies included in this review, only five (2.86% of total) studies are based on semi-supervised methods:

- Helmholtz Principle – [193].
- Information Theory – [207].
- Positive and Unlabeled Learning – [228, 63].
- Pseudo Labels – [155].

### 3.4 Experiment Settings Analysis

The studies included in this review use different combinations of experiment settings, such as different datasets, classification algorithms, and performance metrics. Due to a large number of studies and the consequently large number of experiment’s settings used, define the ideal setting for a new experiment can be very challenging.

The aim of this section is mapping and summarizing the settings of the experiments that are being used to analyze and compare FS methods for text categorization (Research Question 3). We focus on analyzing the following settings:

- What text representation are being used? (Section 3.4.1)
- What public datasets, language of text corpora in datasets, and dataset domains are being used to evaluate the methods? (Section 3.4.2)
- What classifier algorithms are being used to evaluate the effectiveness of FS methods? (Section 3.4.3)
- Which validation settings are the most used? (Section 3.4.4)

We aim to help the design of new researches by providing a summary of which experiment settings are being used. Additionally, we have identified which settings are desirable and are underutilized.

### 3.4.1 Text Representation Used in Experiments

Textual data can be represented in different formats for text classification. In Chapter 2, we present the widely used  $N$ -gram and Word Embedding representation models. Considering the works included in this review, Table 3.6 shows that 88.57% of the methods were evaluated using exclusively Bag of Words (BoW) (uni-gram). Among the remaining works, no other mode of representation was found to be prevalent. It is interesting to note that eight studies used the combination of different representations, two of which combining BoW and Word Embedding.

### 3.4.2 Datasets Used in Experiments

The primary way to evaluate the effectiveness/efficiency of a FS method is training and measuring the performance of a classifier using the FS method. In this section, we indicate which public datasets are most commonly used in FS studies for text classification. We also map the most frequently used languages and domains.

**Public Datasets** Most of the papers included in this review used public datasets to evaluate the proposed methods. Few studies have used private or specifically collected datasets. The use of public datasets is recommended because it facilitates the comparison of methods. Table 3.7 presents the most used public datasets. As our review mapped a considerable number of studies and each one can use several different datasets, a list of

all datasets would be very long and would mostly include datasets that were used by a single study. For this reason, we focus on mapping and presenting in Table 3.7 only the datasets that are public and that were used by at least two studies mapped in our review.

**Language of Text Corpora in Datasets** The majority of the papers included in this review (72.57%) used only English text corpora to evaluate their FS methods. The second most used language is Chinese (26 studies). The third language is Arabic (seven studies). Only four studies perform experiments using two different languages (English and Chinese). Wang and Hong [202] were the only ones who used three different datasets languages (English, Turkish and Kurdish Sorani) in the same study. Table 3.8 presents the languages with at least two studies utilizing that language in their datasets.

The following languages were considered by only one study: Serbian [127], Hinglish [151], Indian [192], Tibetan [72], Vietnamese [62], Japanese [50], Russian [198], Indonesian [173],

Table 3.6: Text representation models to evaluate FS methods.

Representation	Number of Studies	Example References
Bag of Words (BoW) (Uni-gram)	155	[60] [194] [153]
BoW (Uni-gram) + Part of Speech (POS)	2	[72] [149]
BoW (Uni-gram) + Termset	1	[13]
BoW (Uni-gram) + Word Embeddings	2	[92] [233]
$N$ -gram ( $N > 1$ )	1	[2]
$N$ -gram + Part of Speech (POS)	1	[224]
POS + Chunk based Features	1	[196]
POS + Lexicon + Word Embeddings	1	[179]
POS-Pattern (3-gram)	1	[223]
Word Embeddings	2	[189] [93]
Bag of Discriminative Words (BoDW)	1	[234]
Dense word co-occurrence matrix	1	[200]
Meta-features	1	[28]
Context specific features	5	[61] [190] [106] [41] [181]
<b>Total</b>	175	-

Table 3.7: Most commonly used public datasets to evaluate FS methods.

Dataset	Domain	Language	Studies	Percentage
Reuters-21578	News	English	57	32.57%
20NewsGroup	News	English	40	22.86%
WebKB	Web Content	English	19	10.86%
Oshumed	Medical	English	12	6.86%
Fudan	News/Web Content	Chinese	8	4.57%
TDT/TDT2	News	English	8	4.57%
TREC	Open Domain Questions	English	8	4.57%
WAP	Web Content	English	7	4.00%
Sogou	News	Chinese	6	3.43%
Sector	Web Content	English	4	2.29%
UCI Datasets	Several domains	English	4	2.29%
Enron	Email (Spam)	English	3	1.71%
k1a/k1b	Web Content	English	3	1.71%
RCV1	News	English	2	1.14%

Italian [44], and Malay [9].

### 3.4.3 Classification Algorithms Used in Experiments

The studies included in this review propose new or improved FS methods for text classification. To evaluate the performance of the proposed method, the authors perform the classification task using one or more classification algorithms. The choice of the classification algorithms for the experiment directly impacts the classification result and, therefore, the evaluation of the proposed method. Table 3.9 and Fig. 3.7 present the most used classification algorithms in studies experiments.

The most used algorithms are NB and SVM because they are recognized as having good results in the task of classifying texts [4]. Table 3.10 presents the distribution of studies by the number of algorithms used.

### 3.4.4 Validation Settings Used in Experiments

When designing a new experiment, the scientists must clearly define the validation method and whether any statistical tests will be performed to refute or not their hypotheses. This

Table 3.8: Most used language of text corpora in datasets to evaluate FS methods.

Language	Studies	Percentage
English	127	72.57%
Chinese	26	14.86%
Arabic	7	4.00%
Persian	2	1.14%
Turkish	2	1.14%
English and Chinese	4	2.29%

Table 3.9: Classifiers that are most often used to evaluate FS methods.

Algorithm	Studies	Percentage
Support Vector Machines (SVM)	103	58.86%
Naive Bayes (NB)	99	56.57%
$k$ -Nearest Neighbors (KNN)	45	25.71%
Decision Tree (DT)	22	12.57%
Random Forest (RF)	11	6.29%

Table 3.10: Number of classifiers used to evaluate FS methods.

Number of Tested Classifiers	Studies	Percentage
1 classifier	89	50.86%
2 classifiers	43	24.57%
3 classifiers	21	12.00%
4 classifiers	15	8.57%
5 or more classifiers	7	4.00%
<b>Total</b>	<b>175</b>	<b>100%</b>

section presents the main evaluation settings used in studies included in this review.

**Validation Method** To evaluate the proposed method, classification algorithms need to be trained and tested using different datasets. This is usually done by:

- Performing  $k$ -fold cross-validation. In cross-validation, the data is partitioned into  $k$  subsets, called folds. The learning algorithm is then applied  $k$  times, each time one different fold is selected as the test set, and the remaining are used as the training set [165].
- Splitting the dataset into two different sets (training and test sets). Some studies use the standard split between training and testing available in some public datasets.

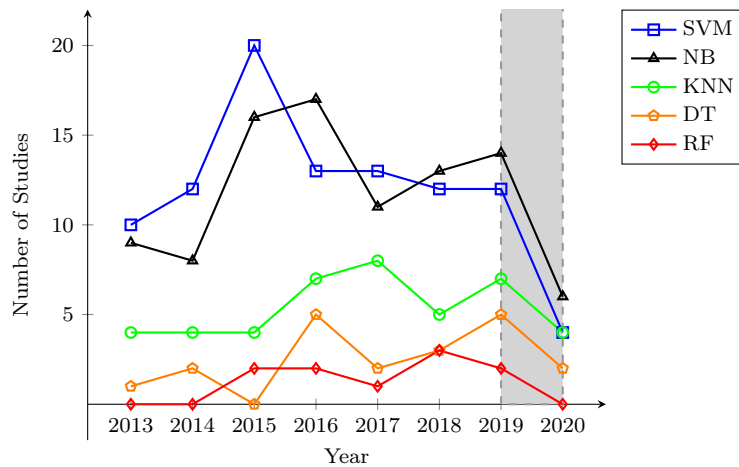


Figure 3.7: Classifiers that have been most often used to evaluate FS methods over the years. The gray box indicates that data collected for 2020 may be preliminary since this review was updated in October 2020.



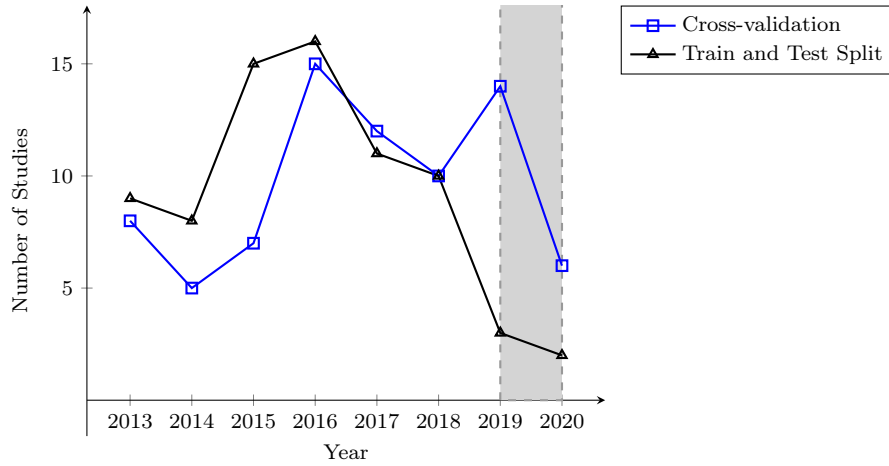


Figure 3.8: Most used validation methods used to evaluate FS methods over the years. The gray box indicates that data collected for 2020 may be preliminary since this review was updated in October 2020.

Other studies define their criteria for this division. The most common is the division based on predefined percentages. However, some studies perform division based on other criteria, such as time division or varying the size of each set within a predefined range.

Approximately half (44.00%) of the studies covered in this review were cross-validated and the other half (43.43%) used different sets of training and testing. Table 3.11 and Fig. 3.8 present the validation methods used.

Table 3.11: Validation methods used in experiments.

Validation Method	Studies	Percentage
10-Fold Cross-validation	55	31.43%
5-Fold Cross-validation	15	8.57%
4-Fold Cross-validation	2	1.14%
3-Fold Cross-validation	4	2.29%
Random Cross-validation	1	0.57%
40% Train + 60% Test	1	0.57%
50% Train + 50% Test	13	7.43%
60% Train + 40% Test	2	1.14%
65% Train + 35% Test	1	0.57%
67% Train + 33% Test	2	1.14%
70% Train + 30% Test	12	6.86%
75% Train + 25% Test	3	1.71%
80% Train + 20% Test	3	1.71%
90% Train + 10% Test	2	1.14%
Dataset Original Split	33	18.86%
Time Split	2	1.14%
Variable Length Training Set	2	1.14%
Not Described	22	12.57%
<b>Total</b>	<b>175</b>	<b>100%</b>

**Statistical Significance Test** The machine learning community has become increasingly aware of the need for statistical validation of the published results [39]. Studies covered in this survey usually evaluate the efficacy of the proposed methods by comparing the proposed solution to other FS methods. The purpose of the comparison is to verify whether the use of the proposed method increases the accuracy/precision/coverage of the classification activity in contrast to the other FS methods. Although practically all studies performed comparisons to demonstrate an improvement in classification performance, we identified that only 29.71% of them used some statistical method to confirm the statistical significance of the results. Table 3.12 shows which statistical methods have been used for this purpose.

Table 3.12: Statistical significance tests used in studies to reject or not the null hypothesis.

Statistical Test	Studies	Percentage
Does not perform any statistical test	123	70.29%
Student's <i>t</i> -Test	31	17.71%
Wilcoxon	7	4.00%
Friedman-test	5	2.86%
Nemenyi	3	1.71%
Analysis of Variance (ANOVA)	2	1.14%
Z-test	2	1.14%
Chi-square	1	0.57%
Cohen's Kappa Statistic	1	0.57%
<b>Total</b>	<b>175</b>	<b>100%</b>

## 3.5 Discussion

Based on the analysis of the problems, methods, and experiment settings raised in this review, we found relevant research trends and discussion points. In this section, we detail these research trends presenting our view on each of them. Sections 3.5.1 to 3.5.4 present research trends and discussions based on each perspective of categorization model that we propose in Section 3.3. Sections 3.5.5 to 3.5.7 present research trends and discussions about experiment settings mapped in Section 3.4.

### 3.5.1 Filter has Been the Feature Selection Dominant Strategy for Text Classification, but a Change is Coming

In Section 3.3.1, we identified that most studies about FS for text classification implement the filter strategy. We found three main reasons for this preference for filter strategy [85, 187]:

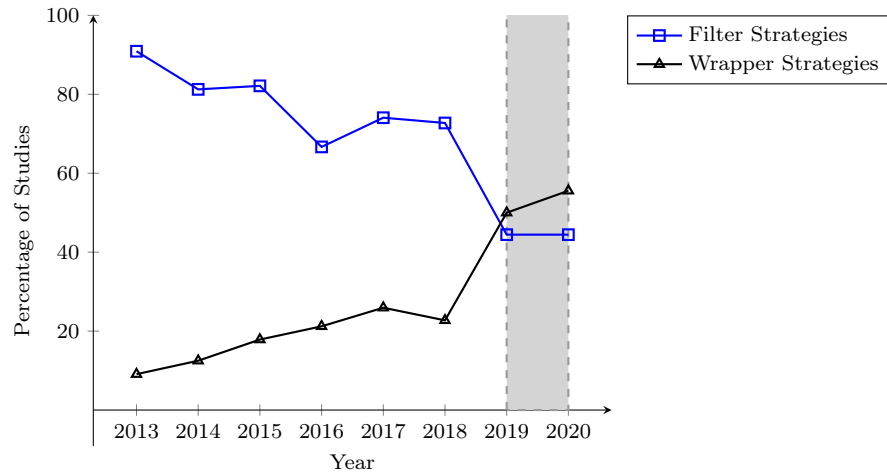


Figure 3.9: Percentage of Filter Strategies vs Percentage of Wrapper Strategies over the years. The gray box indicates that data collected for 2020 may be preliminary since this review was updated in October 2020.

- **Simplicity** – Filter-based methods usually have simpler design and development than wrapper and embedded methods.
- **Classifier independence** – In filter strategy, the result of FS is not biased by choice of classifiers.
- **Computationally efficient** – Filter-based methods are efficient and fast to compute. This advantage becomes even more important for text classification and other problems with high-dimensional data.

Despite filtering be the widely used strategy, we found that the percentage of studies based on filter strategy is decreasing, and the rate of studies based on wrapper strategy is increasing as shown in Table 3.13 and Fig. 3.9. The columns about Filter Strategy encompass the Two-Step (Filter+Filter) studies and the columns about Wrapper Strategy cover Hybrid (Filter+Wrapper) studies. We believe that the percentage of studies using other strategies (wrapper, embedded, and hybrid) will continues to increase. We see the

Table 3.13: Filter Strategies versus Wrapper strategies over the years.

Year	Total of Studies	Filter Strategies		Wrapper Strategies	
		Studies	Percentage	Studies	Percentage
2013	22	20	90.91%	2	9.09%
2014	16	13	81.25%	2	12.50%
2015	28	23	82.14%	5	17.86%
2016	33	22	66.67%	7	21.21%
2017	27	20	74.07%	7	25.93%
2018	22	16	72.73%	5	22.73%
2019	18	8	44.44%	9	50.00%
2020	9	4	44.44%	5	55.56%
<b>Total</b>	<b>175</b>	<b>114</b>	<b>65.14%</b>	<b>39</b>	<b>22.29%</b>

following reasons for this increase:

- Large volume of published studies using the filter strategy – Since the volume of work using the filter strategy is large, we believe that the margin for improvement of results using this strategy is reduced. Therefore, we see that researchers tend to explore other strategies to pursue better results.
- Evolution of computing power and cost – The increase in processing power and computational cost reduction facilitates research techniques that are more computationally intensive, such as wrappers methods. In other words, the computational efficiency of the filter strategy tends to become a less important factor, as the available computing power increases.

### 3.5.2 Metaheuristic Approach is the Trend

In Section 3.3.2, we identified that most studies about FS for text classification are mainly based on the statistic-based approach. However, we have analyzed the evolution of approaches used by grouping them by publication year (Fig. 3.10), and we noticed that the number of studies based on statistical approaches has been decreasing since 2016. On the same graph, we can see a gradual increase in the number of studies based on metaheuristics from 2015 to 2019. In 2016, 4 times more studies were published based on statistical approaches compared to the number of studies based on metaheuristics in the same year. On the other hand, almost the same number of studies were published in each approach during 2019. If this trend continues, in the coming years, the predominant approach will be metaheuristic.

We believe that the increasing use of the metaheuristic approach is motivated by the same factors then the use of the wrapper strategy (discussed in Section 3.5.1). Similarly, a considerable volume of studies is already available based on purely statistical approaches. In this way, researchers tend to have a smaller margin to achieve better results using the same approach. For this reason, they tend to explore more sophisticated approaches, such as metaheuristic techniques. The increasing use of the metaheuristic approach can also be directly related to the wrapper strategy's greater use. Wrapper strategy is frequently employed to search for the best subset of features (as explained in Section 3.2). Since this subset search is a hard problem (as explained in Section 3.2.2), metaheuristic search techniques are usually the solution adopted.

In addition to the two predominant approaches (statistical and metaheuristic), semantic-based and machine learning-based approaches have also been used in a relevant number

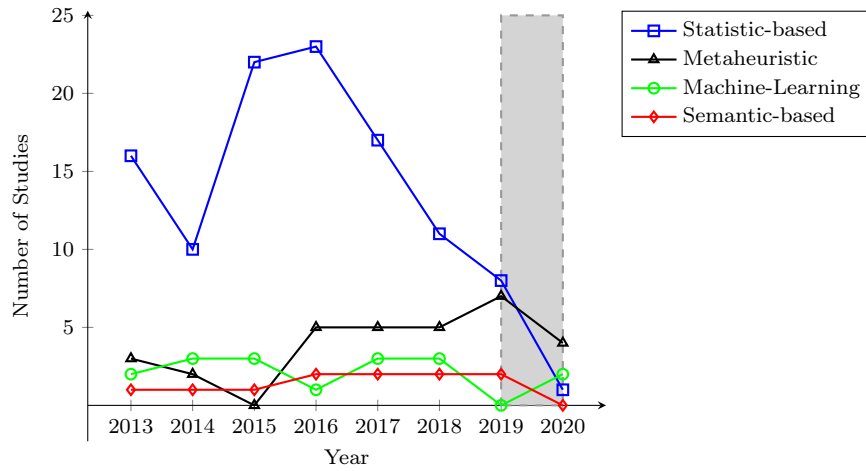


Figure 3.10: Number of FS studies by approach over the years. The gray box indicates that data collected for 2020 may be preliminary since this review was updated in October 2020.

of studies. However, as it is possible to observe in the graph by year (Fig. 3.10), neither approach had a significant increase or decrease in the volume of studies per year. Approaches that have had few scattered studies over the years were not included in this chart. For example, we only found one study mainly based on the grammatical approach published in 2015 and three rule-based studies published in 2016, 2018, and 2019.

Considering the observations made in the previous paragraphs, we conclude that the metaheuristic approach tends to become prevalent in the coming years and the number of studies based mainly on a statistical approach tends to continue decreasing. We also believe that researchers will tend to combine two or more approaches in the same study to seek better results. This review focused on mapping the principal approach used in each study. A future work could be examining all secondary approaches used in each study and how they are being combined.

### 3.5.3 Multiclass Classifiers are Still Dominant

In Section 3.3.3, we identified that most studies were evaluated or designed to multiclass classifiers (67.43% of total) and binary classifiers (24.57% of total). To assess the possible change of this trend, we analyzed the distribution by year (Fig. 3.11). Disregarding the year 2020 with preliminary data, it is not possible to identify any clear trend indicating a change in this distribution. Despite this, we believe that the number of multi-label studies will increase due to the popularity increase of multi-label classification [142]. However, we believe that this type of change tends to be gradual. The reason is that new FS

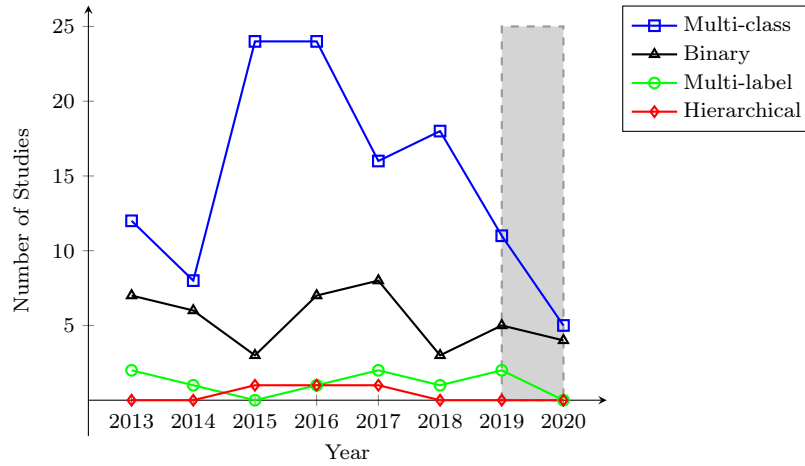


Figure 3.11: Number of FS studies by type of classification over the years. The gray box indicates that data collected for 2020 may be preliminary since this review was updated in October 2020.

studies tend to use the same types of classifiers and datasets that have been widely used in previous studies to facilitate comparison between studies.

### 3.5.4 Supervised Versus Unsupervised Feature Selection Methods

In Section 3.3.4, we identified that most studies are based on supervised techniques (94.86% of total). To assess if there is any sign of growth in studies of unsupervised or semi-supervised techniques, we analyzed this distribution by year (Table 3.14). However, from this table, we can see that the largest number of unsupervised and semi-supervised studies were concentrated in 2017 without progression in the following years. Thus, we believe that for the next few years, studies will probably remain focused on supervised techniques.

Table 3.14: Number of supervised, unsupervised and semi-supervised studies over the years.

Year	Supervised	Unsupervised	Semi-supervised	Percentage of Supervised Studies
2013	21	0	1	95.45%
2014	15	0	1	93.75%
2015	28	0	0	100.00%
2016	29	3	1	87.88%
2017	25	1	1	92.59%
2018	21	1	0	95.45%
2019	18	0	0	100.00%
2020	9	0	0	100.00%
<b>Total</b>	<b>166</b>	<b>5</b>	<b>4</b>	<b>94.86%</b>

Table 3.15: Age and size of most used datasets in experiments.

Dataset	Year of Creation	Number of Documents	Reference
Reuters-21578	1987	21.578	[100]
20NewsGroup	1995	20.000	[156]
WebKB	1997	8.282	[208]

Table 3.16: Historical trends in the usage of content languages for websites [199].

Position	Language	Percentage	Position	Language	Percentage
1	English	54.40%	11	Polish	1.60%
2	Russian	6.70%	12	Chinese	1.60%
3	German	5.30%	13	Dutch, Flemish	1.10%
4	Spanish	4.90%	14	Korean	0.90%
5	French	3.70%	15	Czech	0.90%
6	Japanese	3.40%	16	Vietnamese	0.80%
7	Portuguese	2.70%	17	Arabic	0.70%
8	Italian	2.10%	18	Greek	0.60%
9	Persian	2.10%	19	Hungarian	0.50%
10	Turkish	1.60%	20	Swedish	0.50%

### 3.5.5 Recent Researches Still Over Old Public Datasets: The Need for New Benchmarks

In Section 3.4.2, we present that the three most commonly used datasets in the studies are Reuters-21578, 20NewsGroup, and WebKb. Table 3.15 shows the year of creation and the size (number of documents) each of these datasets. Note that the three datasets are over 20 years old and have a volume below 22,000 documents. These datasets can be considered old and small compared to other datasets like Reuters Corpus Volume I (RCV1). It is a dataset of over 800,000 manually categorized newswire stories made available by Reuters Ltd. for research purposes [101]. Although the RCV1 be a well-known benchmark for text classification with more than 2,000 studies citing the original paper [101], none of the FS studies included in this review employed this dataset in their experiments.

We believe that most current studies still use those same datasets to facilitate a comparison of its results to previous studies. Thus, the use of these datasets tends to be preserved. One way to solve this problem is by using both classical as well as newer/larger datasets in new studies. In this way, it will be possible to compare the results to previous works and evaluate the methods using larger benchmarks.

### 3.5.6 The English Language Dominance

As explained in Section 3.4.2, most studies (72.57%) evaluate techniques only in English datasets. We believe that one of the reasons for this focus on the English language is

Table 3.17: Statistical significance tests used in studies to reject or not the null hypothesis.

Year	Studies	Using Significance Test	Percentage
2013	22	3	13.64%
2014	16	2	12.50%
2015	28	8	28.57%
2016	33	9	27.27%
2017	27	7	25.93%
2018	22	7	31.82%
2019	18	10	55.56%
2020	9	6	66.67%
<b>Total</b>	175	52	29.71%

that about 54.40% of internet web pages is written in this language [199]. The remainder of the web pages is distributed over several other languages, such as Russian, German, and Spanish. However, each one of these languages owns less than 7% of the webpages each [199]. Table 3.16 lists the 20 most widely used languages on the internet as of September 2019.

### 3.5.7 Feature Selection is Already a Mature Field Allowing Statistical Evaluations

As presented in Section 3.4.4, most studies (70.29%) do not perform statistical significance tests to reject or not the null hypothesis. Analyzing the data grouping by year (Table 3.17), we noticed a progressive increase in the use of statistical tests. We believe that this increase indicates a maturing of the research area.

Analyzing publications in conferences and journals separately (Table 3.18), we concluded that the use of statistical tests is widespread in papers published in journals. From 90 articles published in conference, only 11 studies (12.22%) use statistical tests to support their findings. On the other hand, 41 of 85 studies (48.24%) published in journals use statistical tests. In both cases, there is an increase in the use of statistical tests over the years. We believe that this demonstrates an increase in maturity in FS studies for text classification.



Table 3.18: Statistical significance tests used in studies to reject or not the null hypothesis (Conference Studies versus Journal Studies).

Year	Conference Studies			Journal Studies		
	Studies	Using Statistical Tests	Percentage	Studies	Using Statistical Tests	Percentage
2013	14	0	0.00%	8	3	37.50%
2014	10	0	0.00%	6	2	33.33%
2015	18	3	16.67%	10	5	50.00%
2016	18	3	16.67%	15	6	40.00%
2017	13	2	15.38%	14	5	35.71%
2018	9	1	11.11%	13	6	46.15%
2019	8	2	25.00%	10	8	80.00%
2020	0	0	0.00%	9	6	66.67%
<b>Total</b>	90	11	12.22%	85	41	48.24%

## Chapter 4

# Crowd-based Feature Selection Method for Text Classification (CrowdFS)

With our SLR, we identified that virtually all the mapped FS methods have a great dependence on the volume of labeled training data. However, the available labeled data set can be limited in many situations, which can degrade the effectiveness of these methods. A number of studies have already proven that aggregating the judgment of several individuals may result in estimates that are close to the real value in different domains, a phenomenon of collective intelligence known as wisdom of the crowds (WoC) [182]. The popularization of crowd-sourcing platforms and initiatives such as Amazon Mechanical Turk and Appen allow for an easy access to WoC. For this reason, this thesis introduces an approach based on collective intelligence to support FS for text classification. Two central issues are related to this thesis statement:

- **Improve Supervised FS Methods** - Apply collective intelligence to improve FS supervised methods that are dependent from labeled data.
- **Improve Unsupervised FS Methods** - Apply collective intelligence to improve FS unsupervised methods that are independent from labeled data.

I propose the CrowdFS (Crowd-based Feature Selection) method composed of two filter FS stages to explore these two central issues. In the first stage, existing FS methods can be used to perform the first filter of features. The second stage uses a collaborative evaluation to achieve the final selection of features. The CrowdFS can be used as a hybrid FS method (Supervised + Unsupervised) using a supervised method in the first stage. However, it can also be used as a fully unsupervised FS method using an unsupervised FS method in the first stage. We performed two different experiments to evaluate the CrowdFS addressing each central issue. In the first experiment, we evaluated the

CrowdFS using a supervised method, and the second experiment assessed the CrowdFS in a fully unsupervised configuration. Performing two different experiments also allowed us to evaluate the CrowdFS in different settings and contexts. The CrowdFS approach is detailed as presented in Figure 4.1.

The initial feature set is composed of terms (words) extracted from all documents. Preprocessing techniques (e.g. stop words elimination and stemming) can be implemented to reduce the initial feature set size [197]. Even so, depending on the size and volume of the training set, the resulting feature set can be considerably large. For this reason, the approach is composed by an automatic filtering first step. The percent of features that will be filtered in each stage (automatic and collaborative) should be adjusted accordingly to the size of the initial set, the collaborative filtering capacity and the final size required.

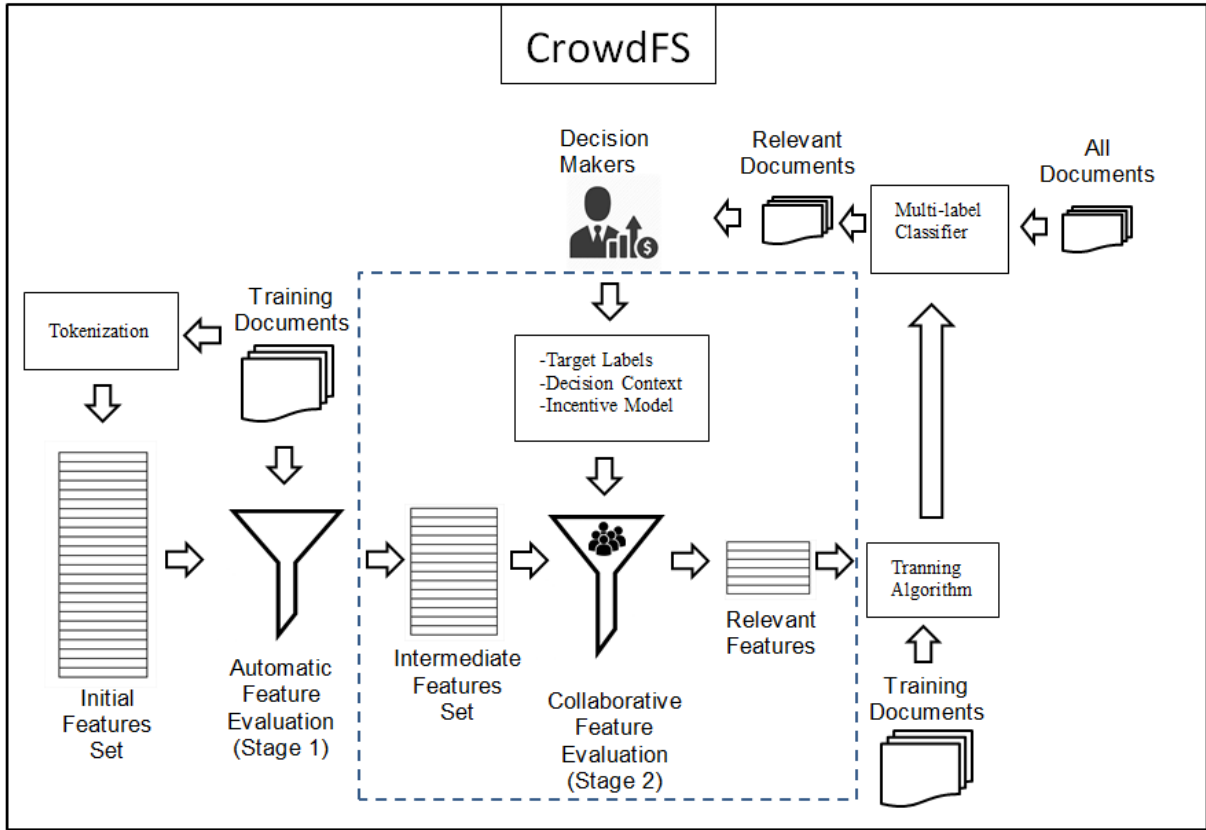


Figure 4.1: Crowd-based Feature Selection (CrowdFS)

Our approach is based on the power of the human crowd work, even not qualified people. Collective intelligence can be defined as a collective decision capability that is at least as good as or better than any single member of the group [65]. Several studies are being conducted in open innovation platforms domain to support the process of selecting the best ideas from a large set [164, 125, 14].

A multi-voting approach [78] known as Bag of Stars (BOS) is commonly adopted in this context of open innovation platforms [81]. In this approach, users are asked to distribute a pre-defined number of  $N$  votes to the best ideas [21]. However a recent study [81] concludes that a multi-vote approach which asks users to distribute votes on the worst ideas, called the Bag of Lemons (BOL), can result in a faster process of collectively filtering ideas. The basic rationale behind this approach is that crowds are much better at eliminating bad ideas than selecting good ones [81].

The problem of filtering best features/terms for each label can be compared to the problem of filtering best ideas. Therefore, the BOS and BOL approaches fit well the goal of intelligent feature selection. From what has been presented in related work we notice the absence of a feature selection method that can be successfully applied in small training set scenarios leading to our proposal presented in next section.

The proposed approach is based on two main hypotheses. The first one is that the collaborative combined evaluation of terms can approach the expert average evaluation. If this is proven, we will know that we can use people considered non-specialists to increase the scalability of the proposed approach. The second hypothesis is that if we use a two-step filtering, including as second step a collaborative selection through voting, we can obtain a classifier with better precision and recall metrics. The Chapters 5 and 6 details the design of the experiment conducted in order to evaluate the feasibility of the CrowdFS approach and test these two associated hypotheses.

# Chapter 5

## First Experiment (Supervised CrowdFS)

Companies, which operate in highly competitive and dynamic markets, need to continuously search for information, inside and outside the company. Information, such as the competitors moves or market regulation's changes, leads to insights and perceptions of opportunities, guiding decision-makers in their choices. Large companies have a team of experts with the task of searching the Internet and other media, for news that might impact a specified decision setting. It is an overwhelming task, since new information keeps continuously feeding the web. To aggravate this scenario, the same piece of information contained in a document might serve to different decision settings (multi-labeling). So, tagging a document once does not mean tagging it forever. To improve the information retrieval, each news article could be classified simultaneously on several aspects or categories, such as the relevance, the level and type of impact for the company, according to related rival companies and the areas of the company that can be affected. However, the task of manually analyzing and classifying each available information item according to all different perspectives is usually not feasible. For this reason, there is a call for effective automatic multi-label text classifiers to support the process of retrieving relevant information for leveraging decision-making scenarios<sup>1</sup>.

### 5.1 Experiment Description

The public dataset Reuters-21578 was chosen for the experiment because it is the main benchmark for evaluation of text classification models[119]. The ModApte Split subset, which is composed of 7,770 training and 3,019 testing documents, was selected because

---

<sup>1</sup>The CrowdFS method and this first experiment was submitted and presented in 21st International Conference Knowledge-Based and Intelligent Information & Engineering Systems (KES)[144].

it is the primary Reuters-21578 subset used for research purposes and because it consists only of documents that have been systematically viewed and evaluated [119].

For this experiment, the categories *Acquisitions*, *Money* and *Oil(Crude)* were chosen because they are among the five categories that have the most documents labeled in the dataset and because they are related to activities realized by the company where the experiment was conducted, a multinational energy company. The number of documents classified for each category in the training and test set is presented in the following table:

Table 5.1: Number of training and testing documents of selected labels.

Label	Training documents	Testing documents
Acquisitions	1650	719
Money	538	179
Oil (Crude)	389	189

The initial features set was obtained by extracting all words from the documents (tokenization) and removing the stop words. We decided not to use other pre-processing methods like stemming and lemmatization to use the original words in the collaborative assessment. For each selected category, an automatic feature evaluation method was performed and the 100 most relevant features of each category were selected. The Chi-square (CS) evaluation method was chosen because it is popular [119] and reached good results when compared to other feature evaluation methods in several text classification benchmarks [220, 176].

The experiment was conducted in the context of information retrieval to feed decision-makers of a Brazilian multinational energy company in a very competitive environment. Subjects were mainly the company,s employees from information technology and exploration & production departments. We used two different multi-voting approaches: Bag of Stars (BOS) and Bag of Lemons (BOL) [81] during the second stage of our feature selection approach. The first strategy, BOS, requested participants to allocate their choices among the features that best identify a document considering a label and the second strategy, BOL, requested participants to select the features that should be discarded, considering the label. For each label 10 positive points (stars to select the best terms) and 10 negative points (lemons to select the worst terms) were provided to participants. In both cases, they were allowed to allocate more than one vote for the same term revealing their degree of certainty in their choices.

Before performing the actual experiment, a pilot study was conducted with three

people in order to validate the design. During the pilot study, we found out that the number of terms presented for each label, coming from the first automatic phase, was excessive (100 words). Participants complained and we realized there were too many, so we decided to set a limit of 50 terms. For the final experiment, we presented only the 50 most relevant words, selected by the automatic phase, per category.

We passed an online questionnaire using the SurveyMonkey tool to obtain the participants profile, as presented in Figure 5.1. The actual experiment involved 29 employees, of which three were experts in selecting documents that would impact decision-makers in competitive scenarios.

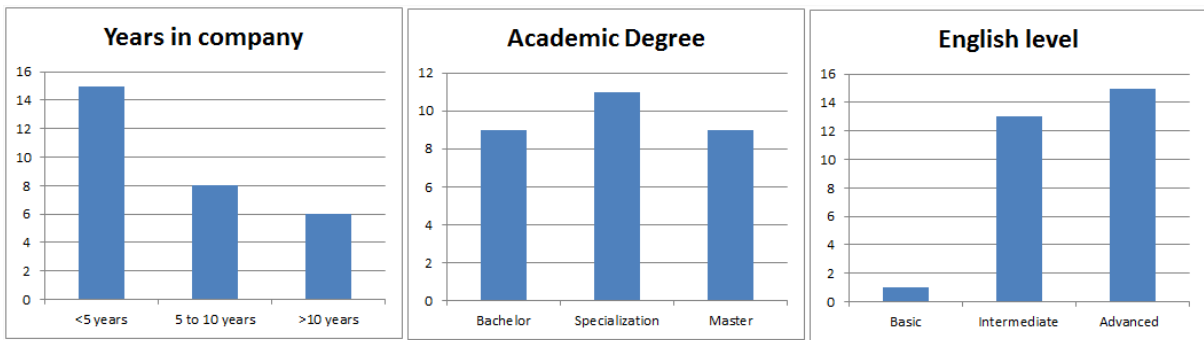


Figure 5.1: Demographics of experiment participants

## 5.2 Data Analysis

The experimental data was collected from all the 29 participants as a crowd. However the analysis has two parts. Firstly, as described in subsection 5.1, we show the crowd participation reached expert average levels. Secondly, as described in subsection 5.2, we show the collaborative approach produces more accurate results, also with better recall than an automatic multi-label text classifier model.

## 5.3 Comparison of the average evaluation of experts and other participants

In order to evaluate the distance between non-specialists and experts evaluation, the evaluations were separated into two groups: specialists and non-specialists. For each group, a ranking of terms was calculated by summing all the evaluations. Table 5.2 shows the number of terms that are at the same time in both rankings considering top 25, 10 and 5 terms with the highest score.

Table 5.2: The intersection between the average evaluation of the specialists compared with an average evaluation of the other participants.

Label	Method	25 top features		10 top features		5 top features	
		Intersect.	%	Intersect.	%	Intersect.	%
Oil(Crude)	BOS	17	68%	8	80%	5	100%
Oil(Crude)	BOL	18	72%	7	70%	4	80%
Acquisition	BOS	15	60%	7	70%	3	60%
Acquisition	BOL	19	76%	7	70%	2	40%
Money	BOS	18	72%	9	90%	3	60%
Money	BOL	15	60%	6	60%	4	80%

This analysis shows that the average of the evaluation performed by people considered non-specialists was close to the average evaluation of the specialists. For example, in a selection of 5 best terms (features) based on the evaluation of the non-specialist participants was on average 70% similar to the top 5 features considering the average evaluation of the experts.

## 5.4 Comparison of the collaborative approach with the automatic approach

The primary objective of this section is to compare the proposed approach (using BOL and BOS) with the selected automatic approach (Chi-square). For this, three rank groups were created according to the term evaluation method, a ranking group based on the sum of the positive evaluations (BOS), another group based on the negative evaluations (BOL) and the last control group with the rankings automatically generated by the Chi-square algorithm (CHI) for each category. Each ranking group is composed of three rankings, one ranking for each category (label). The BOS and Chi-square rankings have the feature relevance proportional to the score, the higher the score, the higher the term rank. The BOL method has the ranking inverted, the more points (votes) a term received the lesser the term rank and relevance.

To compare the learning effectiveness of each evaluation method, a first-order multi-label classifier was developed that, besides the usual parameters (terms matrix and training target label matrix), also receives as parameter which feature set should be considered for each label. A first-order multi-label classifier addresses the problem of multi-label classification by decomposing it into a number of independent binary classification



problems.[227] For this study, the multi-label classifier was developed by the composition of binary classification models based on Support Vector Machines (SVM). This method was selected because it is popular and presents good results when applied to text categorization tasks.[75] The implementation of the multi-label classifier was performed on the MATLAB platform and LIBSVM open library [31].

The comparison between methods was conducted using precision and recall metrics, the two primary metrics used in information retrieval area. [119] Because precision and coverage are inversely proportional metrics, we focus our analysis also on the comparison of the F-measure metric which is the harmonic mean of precision and coverage. In multi-label problems, the effectiveness of the classifier needs to be evaluated considering all labels. For this reason, an averaged version of precision and recall metrics has to be calculated to consider all labels. There are two main ways to compute this average [227]:

- Macro-averaging recall/precision - arithmetic mean of recall/precision metrics computed separately for each label.
- Micro-averaging recall/precision - recall/precision computed from a confusion matrix where each element is obtained by summing the corresponding elements of the confusion matrices of all labels

Using the Macro-averaging, each label will represent exactly the same weight on final averaged value. Using the micro averaging, labels may have different influence in final value depending on the number of positive classified instances of each label. A Macro-averaging version of F-measure can be computed using Macro Averaging precision and recall as follows:

$$\text{Macro-averaging F-measure} = \frac{2 \times \text{Macro-averaging Precision} \times \text{Macro-averaging Recall}}{\text{Macro-averaging Precision} + \text{Macro-averaging Recall}}$$

To analyze the influence of the size of training set on effectiveness of each method, the comparison between methods was performed in two stages. First we performed an analysis considering the complete set of training documents, and then another analysis was performed simulating small training subgroups as follows in the subsections below.

#### 5.4.1 Analysis considering whole training set

For each evaluation method (Chi-Square, BOL and BOS), 50 classification models were trained and tested considering the 50 possible different cut points in the original set of 50 terms selected for each category. Each cut point considers a different number of discarded

terms, with the minimum cut off a single term and the maximum of 49 terms discarded. In this way, it is possible to compare the performance of the evaluation methods in each cut scenario. In this section, training and testing were performed using exactly the standard split between training and testing of the dataset used (ModApte Split).

The comparison of the recall, precision and F-measure results are presented in figs. 3-5.

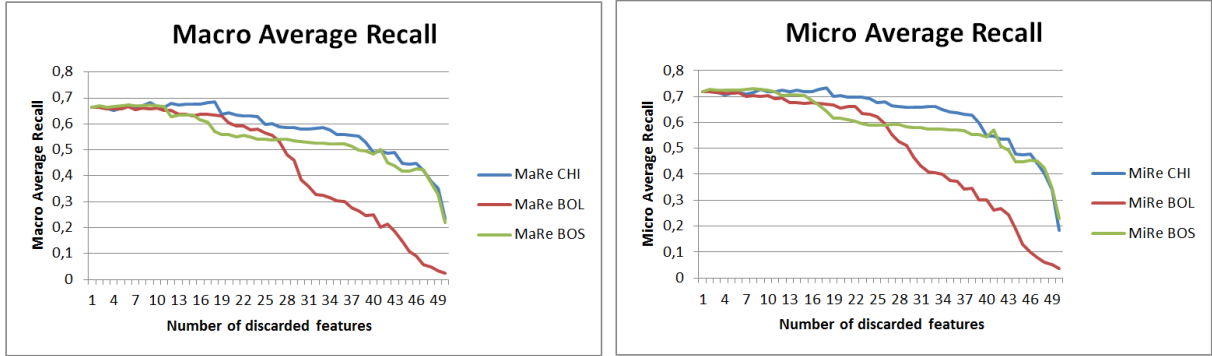


Figure 5.2: Average recall (macro and micro) by evaluation method

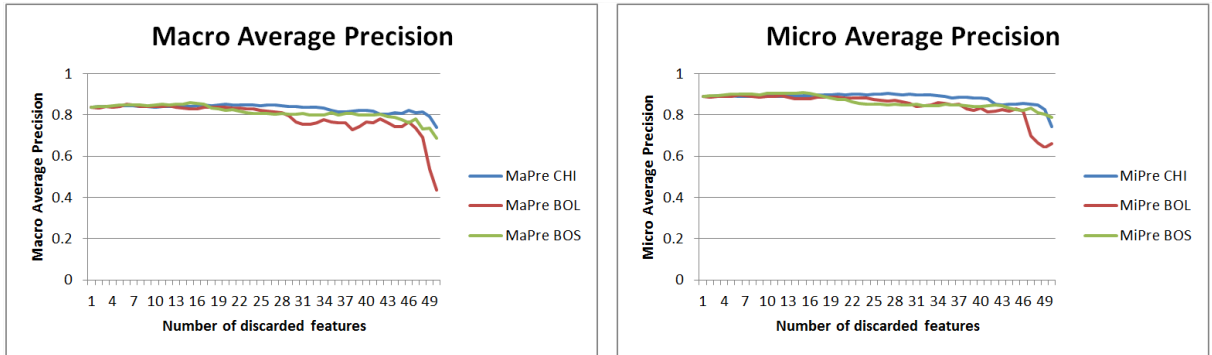


Figure 5.3: Average precision (macro and micro) by evaluation method

For these graphs, the X-axis represents the number of terms discarded for each cut point and the Y-axis represents the value obtained in the metric that is being evaluated considering that cut point. It is possible to note a performance decrease of BOL method when the cut off features exceeded 25 terms, because the ranking generated by BOL focuses on the worst terms. Thus, the best terms for each label have a tied score (no lemon) and therefore there is no rank differentiation between them. For this reason, we can notice that the BOL method tends to achieve better results when a small or medium cut (up to 50%) is desired. For the equivalent reason, the BOS method tends to achieve better results when a more aggressive cut (more than 50%) in the number of terms is performed. After all, using the BOS method, a set of worst terms tends to be tied with no stars.

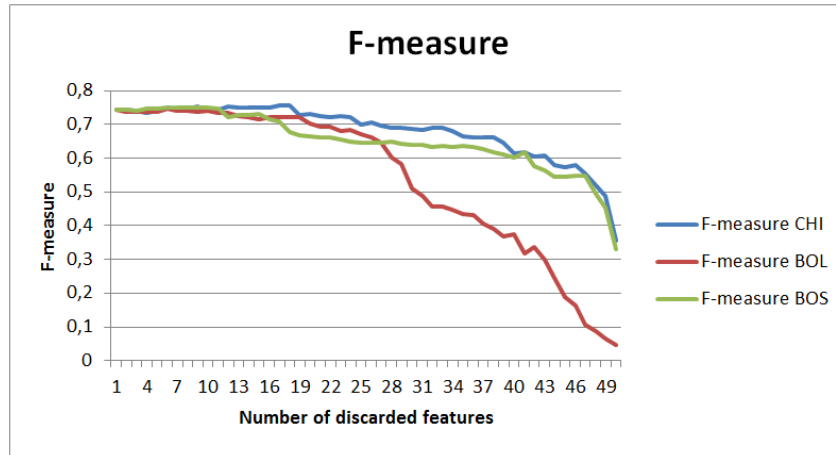


Figure 5.4: Macro averaged F-measure by evaluation method

Analyzing these graphs it is possible to note that neither the collaborative ranking of better terms (BOS) nor the ranking of worse terms (BOL) obtained consistent better results when compared to the automatic method (CHI) when the whole training set was used. As explained in section 4.1, the categories selected in the dataset (Oil, Acquisitions and Money) have a considerable number of training documents. Since the automatic method used (Chi-square) is calculated based on training set examples, it tends to obtain better results when it has a large volume of training instances. For this reason, we conjectured that the CrowdFS approach, by using human common sense, could be more appropriate for small training sets scenarios, precisely when automatic methods tend to get worse results. The next section evaluates the CrowdFS performance in such scenarios by simulating several small training sets extracted from original set.

### 5.4.2 Analysis simulating small training sets

To analyze the CrowdFS performance in small training set scenarios, the original dataset was randomly divided into 100 different training subsets, each one composed of 77 documents. As each subset is composed by different documents, the initial set of terms (features) can be different for each subset. As it was not viable to repeat the collaborative part of the experiment 100 times, we use the original answers to simulate the collaborative evaluations in each subset scenario as explained below.

The 25 top ranked terms using BOS method were extracted and stored. For each subset, the automatic Chi-square ranking was extracted and the 200 top ranked terms were stored. For each subset, this ranking was adjusted by moving all BOS top ranked terms to the top of the ranking keeping the BOS original relative ranking between these

terms. We used only the BOS ranking here, because this part of the analysis simulates aggressive feature cut scenarios (5, 10, 15, 20 and 25 remaining features).

For each training subset in each cut scenario, two models were trained and tested. A first model based on original Chi-square ranking and a second model based on the BOS adjusted ranking. The recall and precision (macro and micro average) comparisons between both models are presented in tables 5.3 and 5.4.

Table 5.3: Comparison of CHI and BOS recall and precision results.

Num. of Features	Ranking	Macro Averaged Recall	Macro Averaged Precision	Micro Averaged Recall	Micro Averaged Precision	Macro Averaged F-measure
5	CHI	0.206 $\pm$ 0.078	0.656 $\pm$ 0.145	0.250 $\pm$ 0.094	0.768 $\pm$ 0.083	0.305 $\pm$ 0.099
5	<b>BOS</b>	<b>0.240 <math>\pm</math> 0.099</b>	<b>0.662 <math>\pm</math> 0.155</b>	<b>0.276 <math>\pm</math> 0.116</b>	<b>0.783 <math>\pm</math> 0.069</b>	<b>0.342 <math>\pm</math> 0.115</b>
10	CHI	0.240 $\pm$ 0.091	0.659 $\pm$ 0.129	0.306 $\pm$ 0.107	0.776 $\pm$ 0.075	0.342 $\pm$ 0.103
10	<b>BOS</b>	<b>0.280 <math>\pm</math> 0.088</b>	<b>0.695 <math>\pm</math> 0.118</b>	<b>0.344 <math>\pm</math> 0.101</b>	<b>0.789 <math>\pm</math> 0.046</b>	<b>0.391 <math>\pm</math> 0.096</b>
15	CHI	0.262 $\pm$ 0.09	0.651 $\pm$ 0.131	0.334 $\pm$ 0.103	0.770 $\pm$ 0.073	0.365 $\pm$ 0.096
15	<b>BOS</b>	<b>0.286 <math>\pm</math> 0.091</b>	<b>0.666 <math>\pm</math> 0.125</b>	<b>0.352 <math>\pm</math> 0.106</b>	<b>0.778 <math>\pm</math> 0.064</b>	<b>0.392 <math>\pm</math> 0.099</b>
20	CHI	0.274 $\pm$ 0.093	0.640 $\pm$ 0.141	0.347 $\pm$ 0.108	0.761 $\pm$ 0.078	0.374 $\pm$ 0.099
20	<b>BOS</b>	<b>0.295 <math>\pm</math> 0.097</b>	<b>0.661 <math>\pm</math> 0.134</b>	<b>0.364 <math>\pm</math> 0.108</b>	<b>0.768 <math>\pm</math> 0.069</b>	<b>0.399 <math>\pm</math> 0.102</b>
25	CHI	0.291 $\pm$ 0.098	0.638 $\pm$ 0.135	0.35 $\pm$ 0.108	0.755 $\pm$ 0.076	0.390 $\pm$ 0.101
25	<b>BOS</b>	<b>0.302 <math>\pm</math> 0.097</b>	<b>0.659 <math>\pm</math> 0.128</b>	<b>0.373 <math>\pm</math> 0.108</b>	<b>0.762 <math>\pm</math> 0.068</b>	<b>0.405 <math>\pm</math> 0.101</b>

Table 5.4: Percentage Increase in Precision, Recall and F-measure metrics.

Num. of Features	Macro Averaged Recall	Macro Averaged Precision	Micro Averaged Recall	Micro Averaged Precision	Macro Averaged F-measure
5	16.51%	0.92%	10.40%	1.95%	12.13%
10	16.67%	5.46%	12.42%	1.68%	14.33%
15	9.16%	2.30%	5.39%	1.04%	7.40%
20	7.66%	3.28%	4.90%	0.92%	6.68%
25	3.78%	3.29%	2.19%	0.93%	3.85%

This comparison shows that the collaborative BOS adjusted ranking resulted in higher precision and coverage models in all features cutting scenarios that were analyzed. Additionally, we realize the two-tailed Student's t-test[178] in order to verify if two distributions of results are significantly different from each other in each scenario. The resultant P-Value of each scenario is presented in table 5.5.

Conventionally, the P-value for statistical significance difference is defined as  $P < 0.05$  [132]. Therefore, it is possible note a relevant difference between Chi-square feature set results and BOL adjusted feature set results mainly in macro and micro average coverage

Table 5.5: P-Value for each group of results.

Num. of Features	Macro Averaged Recall	Macro Averaged Precision	Micro Averaged Recall	Micro Averaged Precision	Macro Averaged F-measure
5	<b>0.00004</b>	0.716561	<b>0.027979</b>	0.157236	<b>0.000124</b>
10	<b>0.00000</b>	<b>0.000867</b>	<b>0.000022</b>	0.055242	<b>0.000000</b>
15	<b>0.00006</b>	0.103782	<b>0.005930</b>	0.137976	<b>0.000003</b>
20	<b>0.00001</b>	<b>0.001823</b>	<b>0.002186</b>	0.159951	<b>0.000000</b>
25	<b>0.00935</b>	<b>0.000540</b>	0.086640	<b>0.029568</b>	<b>0.000059</b>

metric. This analysis indicates that the gain in coverage obtained through the use of collective intelligence in the feature selection process was statistically significant.

## 5.5 Discussion about experimental results

The objective of the experiment and respective results analysis presented in this paper was to compare the performance of feature selection approaches (automatic and CrowdFS) considering different numbers of selected features. A future work would be to evaluate a larger set of features using CrowdFS approach to find the number of selected features that maximizes coverage or precision metrics. One alternative would be to distribute complementary subsets of features to different subgroups of participants in order to not overload any of the participants.

Due to the company and Brazilian government regulations, there were several limitations to the use of financial incentives to employees. For this reason, no financial incentive was used in the present study. Thus, another future work would be to evaluate the proposed approach in an environment where it is possible to use financial incentives or to adopt other types of incentive, for example using gamification techniques. Another alternative is to evaluate the usage of crowdsourcing tools such as Amazon Mechanical Turk which already includes financial incentive as part of the process.

# Chapter 6

## Second Experiment (Unsupervised CrowdFS)

Based on the analysis and discussion of the first experiment's results described in the previous chapter, we designed a new experiment with a different configuration to evaluate the CrowdFS method. In the first experiment, we used a supervised method (Chi-square) in the first stage of CrowdFS. In this second experiment, we used an unsupervised method (TFIDF) for this purpose. In this way, we were able to evaluate the CrowdFS in a fully unsupervised configuration. Besides that, this second experiment has the following differences or improvements compared to the first one:

- **Context and execution of the experiment.** We conducted the first experiment inside a Brazilian multinational energy company, where the participants were their employees. Differently, we conducted the second experiment using an open platform (Appen) available to anyone with access to the internet.
- **The number of participants.** This second experiment had 224 participants in contrast to the first experiment, which had 29 participants.
- **More recent and diversified dataset.** In the first experiment, we used the Reuters21578 dataset created in 1987. In our second experiment, we decide to use the 20Newsgroup dataset because it is eight years more recent and has more diverse subjects in their news.
- **More categories (labels).** In this second experiment, we used six classes of the dataset compared to the first experiment that used three classes.
- **The crowd evaluated a larger number of features.** The first experiment evaluated 50 words (or features) in the collaborative phase of the CrowdFS method. In the second experiment, we increased this number to 400 features/words.

The following sections present the detailed description (Section 6.1), and results analysis (Section 6.2) about this second experiment.

## 6.1 Experiment Description

In this experiment, we use the 20Newsgroup dataset [156]. As presented in section 3, this dataset is among the most used datasets to evaluate methods for selecting features. Among the 20 categories available in the 20Newsgroup dataset, we chose the following six categories for our experiment:

- **sci.med** - Medicine and its related products and regulations.
- **sci.space** - Space, space programs, space related research, etc.
- **soc.religion.christian** - Christianity and related topics.
- **rec.autos** - Automobiles, automotive products and laws.
- **comp.os.ms-windows.misc** - General discussions about Windows issues.
- **comp.sys.ibm.pc.hardware** - XT/AT/EISA hardware, any vendor.

Our decision was based on choosing categories that included different groups of knowledge. However, we included two categories with close subjects (windows and hardware) to assess this situation in our experiment. The initial features set was obtained by extracting all words from the documents (tokenization) and removing the stop words. We decided not to use other pre-processing methods like stemming and lemmatization to use the original words in the collaborative assessment.

We conducted this second experiment using Appen, an open data annotation platform. Appen acquired in 2019 the Figure Eight platform (formerly Crowdfower) that combines over 20 years of expertise running data projects, a global crowd of over one million skilled contractors who speak over 180 languages and dialects in over 130 countries <sup>1</sup>. We performed some preliminary tests on the Appen tool to understand the platform’s operation and define our experiment’s parameters. Based on these initial tests, we determined the following settings for our second experiment:

- **50 words for each participant’s evaluation.** In the preliminary assessment, we found that presenting 100 words generated a high number of low-quality responses. When we adjusted to present 50 words per task, the number of low-quality responses was much lower.
- **Contributor Level 1 (All qualified contributors).** We identified that level 1

---

<sup>1</sup>Figure Eight is now Appen. URL: <https://appen.com/figure-eight-is-now-appen/>

resulted in a quality of responses similar to levels 2 and 3, but with a fast throughput.

- **Geography: 1 Country selected (United States)** We restricted to United States participants because the language and news content of the Dataset are related to this country.
- **Payment per judgment** We initially set the value suggested by the Appen tool for the task: U\$ 0.12 per judgment. Due to a low frequency of responses, we tried to increase the amount paid. However, even with the increase in value, we have not identified a rise in the responses' frequency. This issue is detailed and discussed in section 7.2.

We performed this second experiment in two variations/groups:

- **One category per task (Group 1).** In this variation, for each participant 50 words and only one category were presented. The participant was asked to select the most relevant and least relevant words for this category. The figure 6.1 present a example of this task.
- **All categories by task (Group 2).** In this variation, for each participant 50 words and the six categories were presented. The participant was asked to select the most relevant and least relevant words for each category. The figure 6.2 present a example of this task.

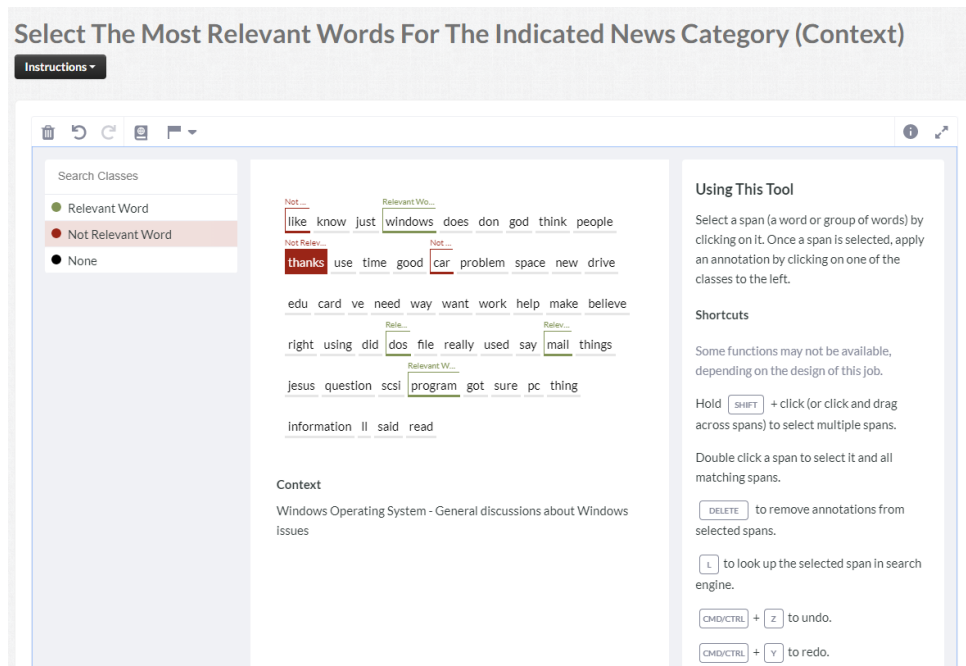


Figure 6.1: Features evaluation functionality developed in Appen and used for the second experiment. Experiment variation: 1 category per task (Group 1).



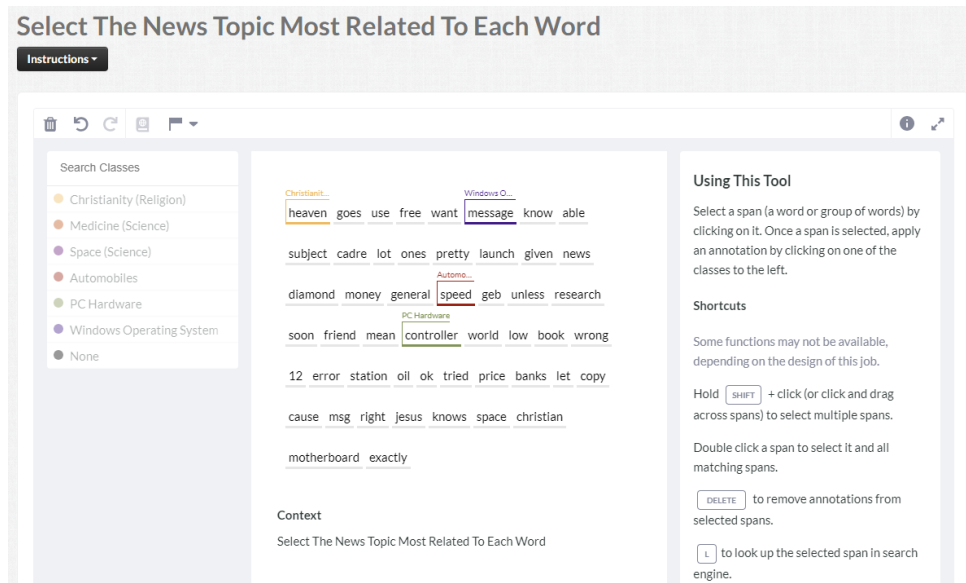


Figure 6.2: Features evaluation functionality developed in Appen and used for the second experiment. Experiment variation: 6 categories per task (Group 2).

## 6.2 Results Analysis

Table 6.1 presents a summary of the number of participants, duration of the collaborative phase, average response time, and financial costs of this second experiment. Discussion points we have identified about the number of participants, response time, and associated costs are detailed in the Chapter 7. In this Results Analysis section, we will focus on evaluating the effectiveness of selecting features using the CrowdFS method in a fully unsupervised configuration.

To assess the effectiveness of CrowdFS, we performed the process of training and testing classification models using the Naive Bayes (NB) algorithm. According to our SLR presented in chapter 3, NB is the most used classifier for feature selection studies for text classification. We use the original training and test of the 20Newsgroups dataset which has a total of 7532 training documents and 11314 test documents. The main site

Table 6.1: Number of participants, duration, and costs for the second experiment.

Experiment Group	Number of task units	Total Number of participants	Number of responses per task unit	Total Duration	Average Time Between Responses	Total Cost	Average Cost per Response
Group 1 (1 category per task)	48	144	3	380.43 h	2.64 h	\$ 30.76	\$ 0.21
Group 2 (6 categories per task)	8	80	10	220.2 h	2.75 h	\$ 9.6	\$ 0.12
<b>Total</b>	-	224	-	600.63 h	2.68 h	\$ 40.36	\$ 0.18

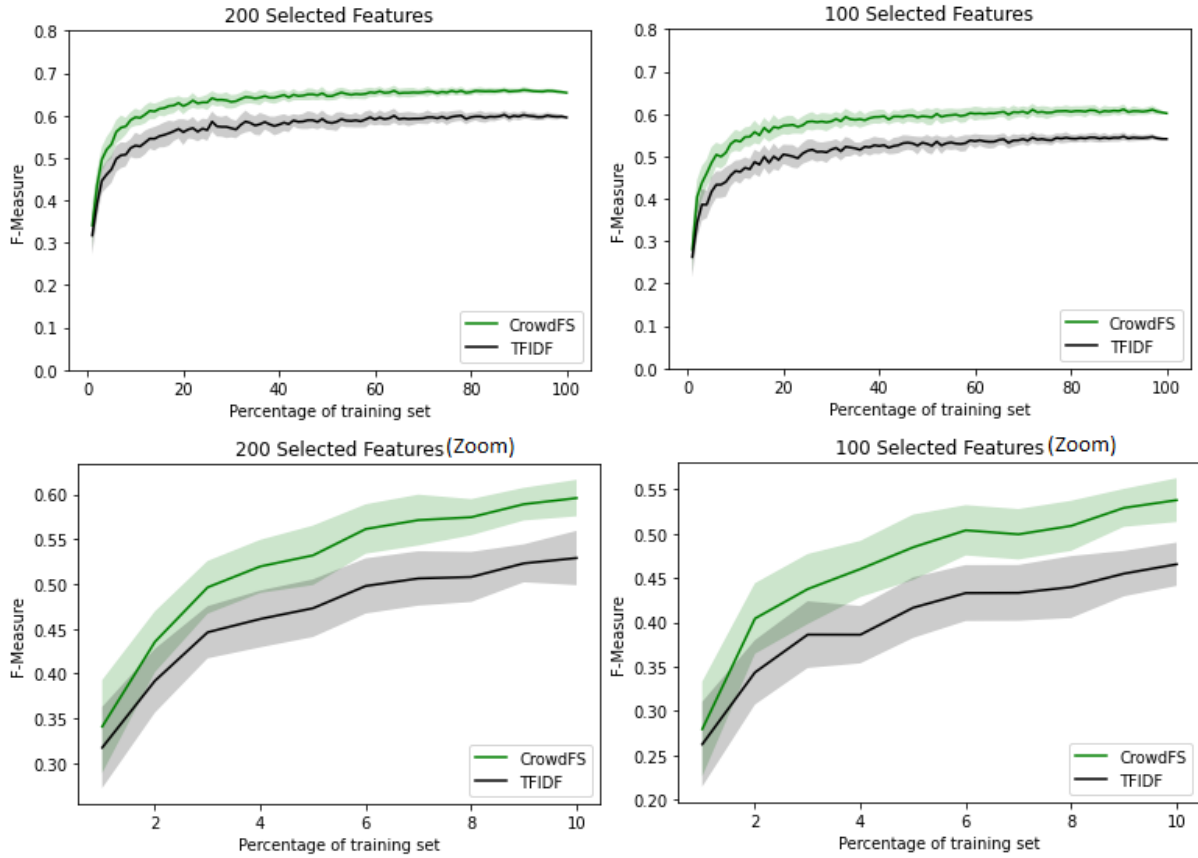


Figure 6.3: Effectiveness comparison between unsupervised CrowdFS and the unsupervised baseline (TFIDF). Experiment variation: 1 category per task (Group 1).

about this dataset<sup>2</sup> recommend the use of this original split instead of performing cross-validation because the train and test sets are separated in time. Considering the 6 classes we chose for the experiment, the total training set was 3561 documents.

We performed the training and testing simulating different sizes of training sets (from 1 % to 100 %) of the total training set (3561 documents). In this way, we were able to compare the CrowdFS method with the baseline (TFIDF) in several different training set sizes. For each training set size (from 1 % to 100 %), we performed the training and test process 30 times. In each repetition, the training documents were selected randomly. Figures 6.3 and 6.4 show the results of this simulation. In these graphs, the lines represent the average F-measure, and the shadow represents the standard deviation based on the 30 results obtained for each training set size. In this second experiment, 400 features were evaluated in the collaborative part of CrowdFS. The graphs show the selection of 50% (200 features) and 25% (100 features) of this total, respectively.

Table 6.2 presents a summary of the information presented in figures 6.3 and 6.4. A

<sup>2</sup>20 Newsgroups. URL: <http://qwone.com/~jason/20Newsgroups/>

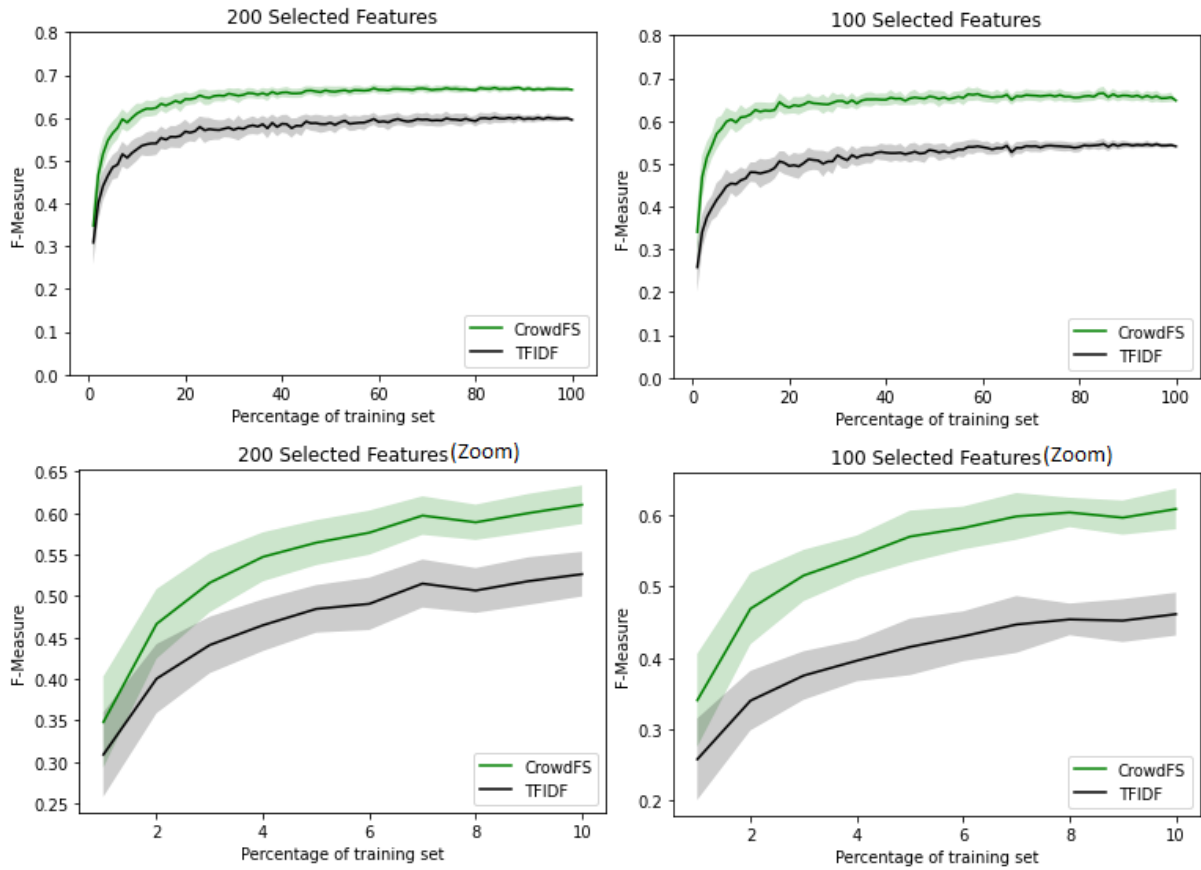


Figure 6.4: Effectiveness comparison between unsupervised CrowdFS and the unsupervised baseline (TFIDF). Experiment variation: 6 categories per task (Group 2).

significant increase in precision and recall (represented by the F-measure metric) can be observed using the CrowdFS method compared to the baseline method (TFIDF).

Table 6.2: Effectiveness comparison between unsupervised CrowdFS and the unsupervised baseline (TFIDF).

Experiment Group	Feature Selection Method	Average F-measure	Average Std. Dev.	Percentage Increase F-measure	Average P-value (CrowdFS x TFIDF)
Group 1 (1 category per task)	CrowdFS (Unsupervised)	57.99	1.68	<b>12.84%</b>	0.00204313
	TFIDF (Baseline)	51.39	1.79	-	-
Group 2 (6 categories per task)	CrowdFS (Unsupervised)	63.95	1.63	<b>24.48%</b>	0.00000327
	TFIDF (Baseline)	51.37	1.82	-	-

## 6.3 Additional Analysis

As an additional analysis, we compared the CrowdFS method in unsupervised mode (using TFIDF in its first stage) with the supervised method used in the first experiment (Chi-square). In the chapter 5, CrowdFS in supervised mode (using Chi-square in its first stage) has already been properly compared with the Chi-square method. However, we consider it interesting to carry out an additional analysis by doing this cross-analysis (Unsupervised CrowdFS versus Supervised Automatic Method). In this comparison, we used the CrowdFS results obtained in the experiment variation of 6 categories per task (Group 2). These additional analysis results are shown in the graphics in figure 6.5.

We used the supervised method in two ways in this comparison:

- **CHI 400** - Using Chi-square to filter features considering only the 400 selected in the first unsupervised stage.
- **CHI** - Using Chi-square to filter features considering all the features available in the Dataset.

The supervised method achieved better results in most of the simulated scenarios. However, in reduced data set scenarios (from 1% to 10% focused on the lower graphs in the figure 6.5 the CrowdFS method in unsupervised mode achieved results similar or superior to the automatic supervised method. This is a very interesting result, since unsupervised methods have the advantage of not being influenced by the volume of labeled data. In the experiment presented in this chapter, we used the TFIDF method for the first stage of CrowdFS. We consider that a relevant future work would be to evaluate the CrowdFS method using other unsupervised methods in its first stage.

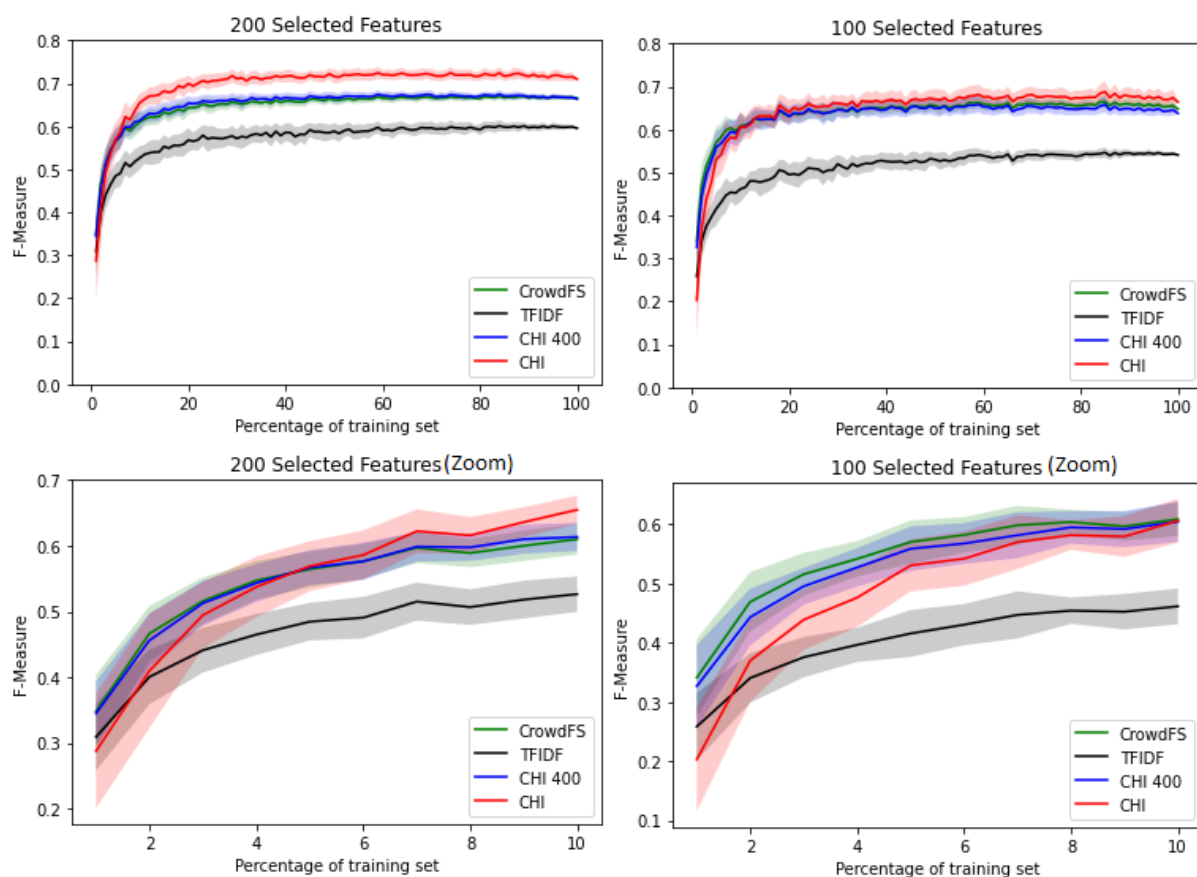


Figure 6.5: Effectiveness comparison between unsupervised CrowdFS and the supervised method (Chi-Square). Experiment variation: 6 categories per task (Group 2).

# Chapter 7

## Conclusion

The volume and diversity of studies included in our SLR show the complexity and importance of FS methods tailored for text classification problems. The categorization scheme for FS methods that we propose in our SLR is an artifact that allows an analysis of the current state of research and the positioning of new studies<sup>1</sup>. This thesis presented the approach called CrowdFS based on collective intelligence techniques to support the FS for text classification in scenarios that are available a small amount of labeled data. The quantitative experiments conducted on a multinational energy company and using an open crowdsourcing platform demonstrated this approach's feasibility, resulting in better accuracy metrics than automatic supervised and unsupervised feature selection methods.

Based on the results and the issues we faced conducting the two experiments described in this thesis, we raised and discussed the main challenges and future studies. The following sections present the main discussion points that we identified applying collective intelligence to support FS for text classification<sup>2</sup>.

### 7.1 Feature Annotation vs Document Annotation

As explained, this thesis's approach focuses on situations where there is a limited set of labeled training data. An alternative to using crowdsourcing to support feature selection would be to use crowdsourcing to categorize unlabeled documents. However, to make available entire documents or excerpts on open platforms may not be possible in many situations due to information security restrictions. An example would be developing

---

<sup>1</sup>The SLR presented in this thesis was submitted and accepted for publication in the journal Artificial Intelligence Review. URL: <https://www.springer.com/journal/10462>

<sup>2</sup>The CrowdFS method and this first experiment was submitted and presented in 21st International Conference Knowledge-Based and Intelligent Information & Engineering Systems (KES)[144].

models within a business environment that involves restricted or confidential documents. For this reason, we consider that making available only a list of words can be a suitable alternative to mitigate this information security problem. In our approach, as only a subgroup of terms is submitted for crowd assessment, people have access to a reduced volume of information, and they do not know the order or combination of words of original documents.

Additionally, we hope that using crowdsourcing to label a word list can be more efficient than evaluating entire documents. However, future work must be conducted to compare each alternative's costs and results to support this hypothesis. Another future work could also examine the collaborative evaluation of features could to support the inference of categories of unlabeled training documents.

## 7.2 Time and cost to use crowd platforms

The frequency of participants' responses using the Appen platform was well below what we were initially expecting (average of 9 responses per day). We did not identify this weakness in our preliminary tests because the first responses tend to be fast. We identified that the slowness occurred after the 30th participant. As a comparison, the first 30 responses took place in 3 hours, while the last 30 responses took place in more than seven days. This low frequency of responses may make it difficult or unfeasible to use this tool for a larger number of participants. We tried to increase the amount paid per unit of work, but it had no significant impact on the responses' frequency. We consider three future configuration changes for a future study to obtain a higher frequency of responses:

- **Allow more than one response per participant.** In our experiment, we set up for each participant to be able to answer only one task unit (50 features) to have a large number of participants. If this setting is changed to allow a person to perform different task units, we believe that the frequency of responses per day can increase considerably.
- **Allow participants from more countries.** As the dataset we used was in English and was created with news from the United States, we restricted our experiment to be able to be performed only by participants located in that country. We believe that if the number of countries is increased in a future experiment, this can also result in a higher frequency of responses.
- **Evaluate another crowdsourcing tool.** A relevant future study would compare

the Appen platform with other crowdsourcing tools, as the Amazon Mechanical Turk. The Appen tool has the advantage of being focused on data annotation to create predictive models. However, other platforms can be evaluated mainly to assess differences in the cost and the average time between responses.

## 7.3 Quality criteria for responses

In our second experiment, in which we used the open crowdsourcing tool Appen, we identified a considerable proportion of responses with patterns indicating low quality. We identified three main patterns:

- None or just one word annotated by the participant.
- All words annotated with the same annotation label.
- Unusual distribution of notes. For example, the first 10 words annotated with the same category, the next 10 words annotated with another category, and so on.

We analyzed these cases individually and identified that the categories selected were not related to the word noted in most cases. We tried to remove the answers with this pattern manually, but this operation did not cause a significant increase in the final result of the feature selection. We believe that the aggregation of responses from several individuals already reduces the impact of individual responses that differ widely from other judgments.

However, in situations where the proportion of low-quality responses is higher or the number of responses for each work item is lower, these low-quality responses can degrade the CrowdFS method's effectiveness. For this reason, we consider the following future studies to address this problem:

- Use test tasks where the predefined expected result is previously known. In this way, it is possible to objectively assess which participants have low quality and disregard their responses.
- Define a specific algorithm or model to identify low-quality responses. This model can use specific patterns per answer or use the distance of each judgment compared to the other participants' answers.
- Implement a positive incentive function to pay more those contributors who submitted quality works.



## 7.4 Number of responses and participants

To define the required number of participants needed to perform the collaborative evaluation of features, we indentified the following points that should be observed:

- **The number of different task units.** The main factor that will define the required crowd's total number of evaluations is the number of task units. The Appen platform suggests that at least three different evaluations for each response.
- **The percentage of low-quality responses.** The higher the percentage of low-quality responses, the greater the number of responses needed to have a good result.

$$\text{Number of Responses} = (\text{Number of Tasks} \times N) + (\text{Number of Tasks} \times A) \quad (7.1)$$

The equation 7.1 represents the number of responses in function of the number of tasks, the number of different responses per unit task ( $N$ ), and the additional number of different responses per unit task ( $A$ ) to compensate for possible losses due to low quality responses.

In addition to this option of pre-defining the number of necessary assessments, the Appen tool provides a dynamic judgment functionality. This option allows each row within the job to collect a variable number of judgments dependent upon contributor's agreement. Future work would be to evaluate the use of this functionality and define the necessary confidence values.

# References

- [1] ABDOLLAHI, M., GAO, X., MEI, Y., GHOSH, S., LI, J. An ontology-based two-stage approach to medical text classification with feature selection by particle swarm optimisation. In *Proceedings of the IEEE Congress on Evolutionary Computation* (2019), p. 119–126.
- [2] AGNIHOTRI, D., VERMA, K., TRIPATHI, P. Computing correlative association of terms for automatic classification of text documents. In *Proceedings of the International Symposium on Computer Vision and the Internet* (2016).
- [3] AGNIHOTRI, D., VERMA, K., TRIPATHI, P. Mutual information using sample variance for text feature selection. In *Proceedings of the International Conference on Communication and Information Processing* (2017), p. 39–44.
- [4] AGNIHOTRI, D., VERMA, K., TRIPATHI, P. Variable global feature selection scheme for automatic classification of text documents. *Expert Systems with Applications* 81 (2017), 268–281.
- [5] AGNIHOTRI, D., VERMA, K., TRIPATHI, P., SINGH, B. Soft voting technique to improve the performance of global filter based feature selection in text corpus. *Applied Intelligence* 49 (2018).
- [6] AGUN, H. V., YILMAZEL, O. Incorporating topic information in a global feature selection schema for authorship attribution. *IEEE Access* 7 (2019), 98522–98529.
- [7] AL-SALEMI, B., AYOB, M., NOAH, S. A. M. Feature ranking for enhancing boosting-based multi-label text categorization. *Expert Systems with Applications* 113 (2018), 531–543.
- [8] AL-SALEMI, B., AYOB, M., NOAH, S. A. M., AZIZ, M. J. A. Feature selection based on supervised topic modeling for boosting-based multi-label text categorization. In *Proceedings of the International Conference on Electrical Engineering and Informatics* (2017), p. 1–6.
- [9] ALSHALABI, H., TIUN, S., OMAR, N., ALBARED, M. Experiments on the use of feature selection and machine learning methods in automatic Malay text categorization. *Procedia Technology* 11 (2013), 748–754.
- [10] ARANI, S. H. S., MOZAFFARI, S. Genetic-based feature selection for spam detection. In *Proceedings of the Iranian Conference on Electrical Engineering* (2013).
- [11] BACCIANELLA, S., ESULI, A., SEBASTIANI, F. Using micro-documents for feature selection: the case of ordinal text classification. *Expert Systems with Applications* (2013).

- [12] BACCIANELLA, S., ESULI, A., SEBASTIANI, F. Feature selection for ordinal text classification. *Neural Computation* (2014).
- [13] BADAWI, D., ALTINCAY, H. A novel framework for termset selection and weighting in binary text classification. *Engineering Applications of Artificial Intelligence* 35 (2014), 38–53.
- [14] BAEZ, M., CONVERTINO, G. Designing a facilitator’s cockpit for an idea management system. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion* (2012), ACM, p. 59–62.
- [15] BAGGENSTOSS, P. M. The PDF projection theorem and the class-specific method. *IEEE Transactions on Signal Processing* 51, 3 (2003), 672–685.
- [16] BAGHERI, A., SARAEE, M., DE JONG, F. Sentiment classification in Persian: introducing a mutual information-based method for feature selection. In *Proceedings of the Iranian Conference on Electrical Engineering* (2013).
- [17] BAHASSINE, S., MADANI, A., AL-SAREM, M., KISSI, M. Feature selection using an improved Chi-square for Arabic text classification. *Journal of King Saud University - Computer and Information Sciences* (2018).
- [18] BAHASSINE, S., MADANI, A., KISSI, M. An improved Chi-square feature selection for Arabic text classification using decision tree. In *Proceedings of the International Conference on Intelligent Systems: Theories and Applications* (2016), p. 1–5.
- [19] BAI, X., GAO, X., XUE, B. Particle swarm optimization based two-stage feature selection in text mining. In *Proceedings of the IEEE Congress on Evolutionary Computation* (2018), p. 1–8.
- [20] BALAKRISHNAMA, S., GANAPATHIRAJU, A. Linear discriminant analysis-a brief tutorial. Relatório Técnico, Institute for Signal and information Processing, 1998.
- [21] BAO, J., SAKAMOTO, Y., NICKERSON, J. Evaluating design solutions using crowds. In *17th Americas Conference on Information Systems 2011, AMCIS 2011* (2011), vol. 5.
- [22] BELAZZOU, M., TOUAHRIA, M., NOUIOUA, F., BRAHIMI, M. An improved sine cosine algorithm to select features for text categorization. *Journal of King Saud University - Computer and Information Sciences* 32, 4 (2020), 454–464.
- [23] BENITEZ, I. P., SISON, A. M., MEDINA, R. P. An improved genetic algorithm for feature selection in the classification of disaster-related Twitter messages. In *Proceedings of the IEEE Symposium on Computer Applications and Industrial Electronics* (2018).
- [24] BENSaid, F., ALIMI, A. M. Online feature selection system for big data classification based on multi-objective automated negotiation. *Pattern Recognition* 110 (2021), 107629.
- [25] BERGSTRA, J., BENGIO, Y. Random search for hyper-parameter optimization. *Journal of Machine Learning Research* (2013).

- [26] BOJANOWSKI, P., GRAVE, E., JOULIN, A., MIKOLOV, T. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.
- [27] BRAYTEE, A., LIU, W., CATCHPOOLE, D., KENNEDY, P. Multi-label feature selection using correlation information. In *Proceedings of the ACM on Conference on Information and Knowledge Management* (2017), p. 1649–1656.
- [28] CANUTO, S., SOUSA, D. X., GONÇALVES, M. A., ROSA, T. C. A thorough evaluation of distance-based meta-features for automated text classification. *IEEE Transactions on Knowledge and Data Engineering* 11, 10 (2018), 346–347.
- [29] CEKIK, R., UYSAL, A. K. A novel filter feature selection method using rough set for short text data. *Expert Systems with Applications* 160 (2020), 113691.
- [30] CHANDRASHEKAR, G., SAHIN, F. A survey on feature selection methods. *Computers and Electrical Engineering* 40, 1 (2014), 16–28.
- [31] CHANG, C.-C., LIN, C.-J. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2, 3 (2011), 27.
- [32] CHEN, H., HOU, Q., HAN, L., HU, Z., YE, Z., ZENG, J., YUAN, J. Distributed text feature selection based on bat algorithm optimization. In *Proceedings of the IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications* (2019), vol. 1, p. 75–80.
- [33] CHEN, H., HOU, Y., LUO, Q., HU, Z., YAN, L. Text feature selection based on water wave optimization algorithm. In *Proceedings of the International Conference on Advanced Computational Intelligence* (2018).
- [34] CHEN, L., LI, J., ZHANG, L. A method of text categorization based on genetic algorithm and LDA. In *Proceedings of the Chinese Control Conference* (2017).
- [35] CHEN, X., MA, J., LU, Y. Feature selection for Chinese online reviews sentiment classification. In *Proceedings of the Joint Conference of International Conference on Computational Problem-Solving and International High Speed Intelligent Communication Forum* (2013).
- [36] CHEN, Y., HAN, B., HOU, P. New feature selection methods based on context similarity for text categorization. In *Proceedings of the International Conference on Fuzzy Systems and Knowledge Discovery* (2014).
- [37] CHOPARD, B., TOMASSINI, M. *An introduction to metaheuristics for optimization*. Springer International Publishing, 2018.
- [38] CHORMUNGE, S., JENA, S. Correlation based feature selection with clustering for high dimensional data. *Journal of Electrical Systems and Information Technology* 5, 3 (2018), 542–549.
- [39] DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research* (2006).

- [40] DENG, X., LI, Y., WENG, J., ZHANG, J. Feature selection for text classification: a review. *Multimedia Tools and Applications* 78, 3 (2019), 3797–3816.
- [41] EKBAL, A., SAHA, S. Joint model for feature selection and parameter optimization coupled with classifier ensemble in chemical mention recognition. *Knowledge-Based Systems* (2015).
- [42] FENG, G., GUO, J., JING, B. Y., SUN, T. Feature subset selection using naive Bayes for text classification. *Pattern Recognition Letters* (2015).
- [43] FENG, L., ZUO, W., WANG, Y. Improved comprehensive measurement feature selection method for text categorization. In *Proceedings of the International Conference on Network and Information Systems for Computers* (2015).
- [44] FERILLI, S., DE CAROLIS, B., ESPOSITO, F., REDAVID, D. Sentiment analysis as a text categorization task: a study on feature and algorithm selection for Italian language. In *Proceedings of the IEEE International Conference on Data Science and Advanced Analytics* (2015).
- [45] FERREIRA, C. H. P., DE MEDEIROS, D. M. R., SANTANA, F. FCFilter: feature selection based on clustering and genetic algorithms. In *Proceedings of the IEEE Congress on Evolutionary Computation* (2016).
- [46] FONG, S., GAO, E., WONG, R. Optimized swarm search-based feature selection for text mining in sentiment analysis. In *Proceedings of the IEEE International Conference on Data Mining Workshop* (2016), p. 1153–1162.
- [47] FORMAN, G. A pitfall and solution in multi-class feature selection for text classification. In *Proceedings of the International Conference on Machine Learning* (2004).
- [48] FRAGOSO, R. C. P., PINHEIRO, R. H. W., CAVALCANTI, G. D. C. Class-dependent feature selection algorithm for text categorization. In *Proceedings of the International Joint Conference on Neural Networks* (2016), vol. 2016-Octob.
- [49] FRAGOSO, R. C. P., PINHEIRO, R. H. W., CAVALCANTI, G. D. C. A method for automatic determination of the feature vector size for text categorization. In *Proceedings of the Brazilian Conference on Intelligent Systems* (2017).
- [50] FUKUMOTO, F., SUZUKI, Y. Temporal-based feature selection and transfer learning for text categorization. In *Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management* (2015).
- [51] GAO, Z., XU, Y., MENG, F., QI, F., LIN, Z. Improved information gain-based feature selection for text categorization. In *Proceedings of the International Conference on Wireless Communications, Vehicular Technology, Information Theory and Aerospace & Electronic Systems* (2014).
- [52] GHAREB, A. S., ABU BAKARA, A., AL-RADAIDEH, Q. A., HAMDAN, A. R. Enhanced filter feature selection methods for Arabic text categorization. *International Journal of Information Retrieval Research* (2018).

- [53] GHAREB, A. S., BAKAR, A. A., HAMDAN, A. R. Hybrid feature selection based on enhanced genetic algorithm for text categorization. *Expert Systems with Applications* (2016).
- [54] GÖKALP, O., TASCI, E., UGUR, A. A novel wrapper feature selection algorithm based on iterated greedy metaheuristic for sentiment classification. *Expert Systems with Applications* 146 (2020), 113176.
- [55] GUNDUZ, H., CATALTEPE, Z. Borsa Istanbul (BIST) daily prediction using financial news and balanced feature selection. *Expert Systems with Applications* (2015).
- [56] GUO, Y., CHUNG, F., LI, G. An ensemble embedded feature selection method for multi-label clinical text classification. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine* (2017).
- [57] GUO, Y., CHUNG, F., LI, G., ZHANG, L. Multi-label bioinformatics data classification with ensemble embedded feature selection. *IEEE Access* 7 (2019), 103863–103875.
- [58] GURU, D., SWARNALATHA, K., KUMAR, V. N., ANAMI, B. Effective technique to reduce the dimension of text data. *International Journal of Computer Vision and Image Processing* 10 (2020), 67–85.
- [59] GURU, D. S., ALI, M., SUHIL, M. A novel term weighting scheme and an approach for classification of agricultural arabic text complaints. In *Proceedings of the IEEE International Workshop on Arabic and Derived Script Analysis and Recognition* (2018), p. 24–28.
- [60] GURU, D. S., SUHIL, M., RAJU, L. N., KUMAR, N. V. An alternative framework for univariate filter based feature selection for text categorization. *Pattern Recognition Letters* 103, 2018 (2018), 23–31.
- [61] HAGENAU, M., LIEBMANN, M., NEUMANN, D. Automated news reading: stock price prediction based on financial news using context-capturing features. *Decision Support Systems* 55, 3 (2013), 685–697.
- [62] HAI, N. T., NGHIA, N. H., LE, T. D., NGUYEN, V. T. A hybrid feature selection method for Vietnamese text classification. In *Proceedings of the IEEE International Conference on Knowledge and Systems Engineering* (2015).
- [63] HAN, J., ZUO, W., LIU, L., XU, Y., PENG, T. Building text classifiers using positive, unlabeled and ‘outdated’ examples. *Concurrency Computation* (2016).
- [64] HIGGINS, J. P. T., GREEN, S. *Cochrane handbook for systematic reviews of interventions: Cochrane book series*. Wiley-Blackwell, 2008.
- [65] HILTZ, S., TUROFF, M. *The Network Nation: Human Communication Via Computer*. MIT Press, 1993.
- [66] HUSSAIN, S., KEUNG, J., KHAN, A. A. Software design patterns classification and selection using text categorization approach. *Applied Soft Computing* 58 (2017), 225–244.

- [67] HUSSAIN, S. F., BABAR, H. Z. U. D., KHALIL, A., JILLANI, R. M., HANIF, M., KHURSHID, K. A fast non-redundant feature selection technique for text data. *IEEE Access* 8 (2020), 181763–181781.
- [68] IMANI, M. B., KEYVANPOUR, M. R., AZMI, R. A novel embedded feature selection method: a comparative study in the application of text categorization. *Applied Artificial Intelligence* (2013).
- [69] ISLAM, M., ANJUM, A., AHSAN, T., WANG, L. Dimensionality reduction for sentiment classification using machine learning classifiers. In *Proceedings of the IEEE Symposium Series on Computational Intelligence* (2019), p. 3097–3103.
- [70] JAPKOWICZ, N. The class imbalance problem: significance and strategies. In *Proceedings of the International Conference on Artificial Intelligence* (2000).
- [71] JAVED, K., MARUF, S., BABRI, H. A. A two-stage Markov blanket based feature selection algorithm for text classification. *Neurocomputing* (2015).
- [72] JIANG, T., YU, H. A novel feature selection based on Tibetan grammar for Tibetan text classification. In *Proceedings of the IEEE International Conference on Software Engineering and Service Sciences* (2015).
- [73] JIANG, X. Y., JIN, S. An improved mutual information-based feature selection algorithm for text classification. In *Proceedings of the International Conference on Intelligent Human-Machine Systems and Cybernetics* (2013).
- [74] JIE, Y., KEPING, L. The fault diagnosis model for railway system based on an improved feature selection method. In *Proceedings of the IEEE International Conference on Electronics Information and Emergency Communication* (2019), p. 1–4.
- [75] JOACHIMS, T. Text categorization with support vector machines: Learning with many relevant features. *Machine learning: ECML-98* (1998), 137–142.
- [76] KARABULUT, M. Fuzzy unordered rule induction algorithm in text categorization on top of geometric particle swarm optimization term selection. *Knowledge-Based Systems* 54 (2013), 288–297.
- [77] KERMANI, F. Z., ESLAMI, E., SADEGHI, F. Global filter-wrapper method based on class-dependent correlation for text classification. *Engineering Applications of Artificial Intelligence* 85 (2019), 619–633.
- [78] KESSLER, F. Team decision making: pitfalls and procedures. *Management Development Review* 8, 5 (1995), 38–40.
- [79] KIM, K., ZZANG, S. Trigonometric comparison measure: a feature selection method for text categorization. *Data & Knowledge Engineering* 119 (2018).
- [80] KITCHENHAM, B. Procedures for performing systematic reviews. Relatório Técnico TR/SE-0401, Department of Computer Science, Keele University and National ICT, 2004.
- [81] KLEIN, M., GARCIA, A. C. B. High-speed idea filtering with the bag of lemons. *Decision Support Systems* 78 (2015), 39–50.

- [82] KOWSARI, K., JAFARI MEIMANDI, K., HEIDARYSAFA, M., MENDU, S., BARNES, L., BROWN, D. Text classification algorithms: a survey. *Information (Switzerland)* 10 (2019).
- [83] KRAMER, M. A. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal* 37, 2 (1991), 233–243.
- [84] KUMAR, H. M. K., HARISH, B. S. Sarcasm classification: a novel approach by using content based feature selection method. *Procedia Computer Science* 143 (2018), 378–386. 8th International Conference on Advances in Computing & Communications (ICACC-2018).
- [85] KUMAR, V. Feature selection: a literature review. *The Smart Computing Review* 4, 3 (2014).
- [86] KUMBHAR, P., MALI, M. A survey on feature selection techniques and classification algorithms for efficient text classification. *International Journal of Science and Research* 14, 5 (2013), 2319–2364.
- [87] KUMBHAR, P., MALI, M., ATIQUE, M. A genetic-fuzzy approach for automatic text categorization. In *Proceedings of the International Advance Computing Conference* (2017).
- [88] KUN, Y. J., LEI, Z. Sentiment feature selection algorithm for Chinese micro-blog. In *Proceedings of the International Conference on Management of e-Commerce and e-Government* (2014), p. 114–118.
- [89] KYAW, K. S., LIMSIRORATANA, S. Towards nature-inspired intelligence search for optimization of multi-dimensional feature selection. In *Proceedings of the International Computer Science and Engineering Conference* (2019), p. 379–384.
- [90] LABANI, M., MORADI, P., AHMADIZAR, F., JALILI, M. A novel multivariate filter method for feature selection in text classification problems. *Engineering Applications of Artificial Intelligence* 70, November 2016 (2018), 25–37.
- [91] LABANI, M., MORADI, P., JALILI, M. A multi-objective genetic algorithm for text feature selection using the relative discriminative criterion. *Expert Systems with Applications* 149 (2020), 113276.
- [92] LAMPOS, V., ZOU, B., COX, I. J. Enhancing feature selection using word embeddings. In *Proceedings of the International Conference on World Wide Web* (2017).
- [93] LAN, Y., HAO, Y., XIA, K., QIAN, B., LI, C. Stacked residual recurrent neural networks with cross-layer attention for text classification. *IEEE Access* 8 (2020), 70401–70410.
- [94] LANDAUER, T. K., FOLTZ, P. W., LAHAM, D. An introduction to latent semantic analysis. *Discourse Processes* 25, 2-3 (1998), 259–284.
- [95] LARABI MARIE-SAINTE, S., ALALYANI, N. Firefly algorithm based feature selection for Arabic text classification. *Journal of King Saud University - Computer and Information Sciences* (2018).



- [96] LAZAR, C., TAMINAU, J., MEGANCK, S., STEENHOFF, D., COLETTA, A., MOLTER, C., DE SCHAETZEN, V., DUQUE, R., BERSINI, H., NOWÉ, A. A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2012).
- [97] LEE, J., KIM, D. W. Feature selection for multi-label classification using multi-variate mutual information. *Pattern Recognition Letters* (2013).
- [98] LEE, J., KIM, D.-W. Mutual information-based multi-label feature selection using interaction information. *Expert Systems with Applications* 42, 4 (2015), 2013–2025.
- [99] LEE, J., YU, I., PARK, J., KIM, D.-W. Memetic feature selection for multi-label text categorization using label frequency difference. *Information Sciences* 485 (2019), 263–280.
- [100] LEWIS, D. D. Reuters-21578 text categorization collection data set, 2019.
- [101] LEWIS, D. D., YANG, Y., ROSE, T. G., LI, F. RCV1: a new benchmark collection for text categorization research. *Journal of Machine Learning Research* 5 (2004), 361–397.
- [102] LI, B. Importance weighted feature selection strategy for text classification. In *Proceedings of the International Conference on Asian Language Processing* (2016).
- [103] LI, B. Selecting features with class based and importance weighted document frequency in text classification. In *Proceedings of the ACM Symposium on Document Engineering* (2016), p. 139–142.
- [104] LI, B., YAN, Q., XU, Z., WANG, G. Weighted document frequency for feature selection in text classification. In *Proceedings of International Conference on Asian Language Processing* (2015).
- [105] LI, J. An approach to meta feature selection. In *Proceedings of the Canadian Conference on Electrical and Computer Engineering* (2013).
- [106] LI, J., ZHAO, J., LU, K. Joint feature selection and structure preservation for domain adaptation. In *Proceedings of the IJCAI International Joint Conference on Artificial Intelligence* (2016).
- [107] LI, L., LI, C. Research and improvement of a spam filter based on naive Bayes. In *Proceedings of the International Conference on Intelligent Human-Machine Systems and Cybernetics* (2015).
- [108] LI, Q., HE, L., LIN, X. Categorical term frequency probability based feature selection for document categorization. In *Proceedings of the International Conference on Soft Computing and Pattern Recognition* (2013).
- [109] LI, Q., HE, L., LIN, X. Dimension reduction based on categorical fuzzy correlation degree for document categorization. In *Proceedings of the IEEE International Conference on Granular Computing* (2013).

- [110] LI, Q., HE, L., LIN, X. Improved categorical distribution difference feature selection for Chinese document categorization. In *Proceedings of the International Conference on Ubiquitous Information Management and Communication* (2014).
- [111] LI, Z., LU, W., SUN, Z., XING, W. A parallel feature selection method study for text classification. *Neural Computing and Applications* 28 (2016), 1–12.
- [112] LIANG, J., ZHOU, X., GUO, L., BAI, S. Feature selection for sentiment classification using matrix factorization. In *Proceedings of the International Conference on World Wide Web* (2015), p. 63–64.
- [113] LIFANG, Y., SIJUN, Q., HUAN, Z. Feature selection algorithm for hierarchical text classification using Kullback-Leibler divergence. In *Proceedings of the IEEE International Conference on Cloud Computing and Big Data Analysis* (2017).
- [114] LIN, K.-C., ZHANG, K.-Y., HUANG, Y.-H., HUNG, J. C., YEN, N. Feature selection based on an improved cat swarm optimization algorithm for big data classification. *The Journal of Supercomputing* (2016).
- [115] LIU, Y., WANG, Y., FENG, L., ZHU, X. Term frequency combined hybrid feature selection method for spam filtering. *Pattern Analysis and Applications* (2016).
- [116] LU, Y., CHEN, Y. A text feature selection method based on the small world algorithm. *Procedia Computer Science* 107 (2017), 276–284.
- [117] LU, Y., LIANG, M., YE, Z., CAO, L. Improved particle swarm optimization algorithm and its application in text feature selection. *Applied Soft Computing Journal* (2015).
- [118] MALJI, P., SAKHARE, S. Significance of entropy correlation coefficient over symmetric uncertainty on FAST clustering feature selection algorithm. In *Proceedings of International Conference on Intelligent Systems and Control* (2017).
- [119] MANNING, C. D., RAGHAVAN, P., SCHÜTZE, H., OTHERS. *Introduction to information retrieval*, vol. 1. Cambridge university press Cambridge, 2008.
- [120] MANNING, C. D., SCHÜTZE, H., RAGHAVAN, P. *Introduction to information retrieval*, vol. 39. Cambridge University Press Cambridge, 2008.
- [121] MANOCHANDAR, S., PUNNIYAMOORTHY, M. Scaling feature selection method for enhancing the classification performance of support vector machines in text mining. *Computers and Industrial Engineering* 124 (2018), 139–156.
- [122] MÉNDEZ, J. R., NEZ, T. R. C.-Y., RUANO-ORDÁS, D. A new semantic-based feature selection method for spam filtering. *Applied Soft Computing* 76 (2019), 89–104.
- [123] MIKOLOV, T., CHEN, K., CORRADO, G., DEAN, J. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations - Workshop Track Proceedings* (2013).

- [124] MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G. S., DEAN, J. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Advances in Neural Information Processing Systems* (2013), p. 3111–3119.
- [125] MILLER, B., HEMMER, P., STEYVERS, M., LEE, M. D. The wisdom of crowds in rank ordering problems. In *9th International Conference on Cognitive Modeling* (2009).
- [126] MIROŃCZUK, M. M., PROTASIEWICZ, J. A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications* 106 (2018), 36–54.
- [127] MLADENOVIĆ, M., MITROVIĆ, J., KRSTEV, C., VITAS, D. Hybrid sentiment analysis framework for a morphologically rich language. *Journal of Intelligent Information Systems* (2016).
- [128] MOHER, D., LIBERATI, A., TETZLAFF, J., ALTMAN, D. G. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Journal of Clinical Epidemiology* (2009).
- [129] NAG, K., PAL, N. R. A multiobjective genetic programming-based ensemble for simultaneous feature selection and classification. *IEEE Transactions on Cybernetics* (2016).
- [130] NAIK, A., RANGWALA, H. Embedding feature selection for large-scale hierarchical classification. In *Proceedings of the IEEE International Conference on Big Data* (2016).
- [131] NAM, L. N. H., QUOC, H. B. A combined approach for filter feature selection in document classification. In *Proceedings of the International Conference on Tools with Artificial Intelligence* (2016).
- [132] NEYMAN, J., PEARSON, E. S. The testing of statistical hypotheses in relation to probabilities a priori. In *Mathematical Proceedings of the Cambridge Philosophical Society* (1933), vol. 29, Cambridge Univ Press, p. 492–510.
- [133] NOGUEIRA RIOS, T., GAMA BISPO, B. V. Statera: a balanced feature selection method for text classification. In *Proceedings of the Brazilian Conference on Intelligent Systems* (2018), p. 260–265.
- [134] ONAN, A., KORUKOGLU, S. A feature selection model based on genetic rank aggregation for text sentiment classification. *Journal of Information Science* 43, 1 (2017), 25–38.
- [135] ONG, B. Y., GOH, S. W., XU, C. Sparsity adjusted information gain for feature selection in sentiment analysis. In *Proceedings of the IEEE International Conference on Big Data* (2015), p. 2122–2128.
- [136] ORTEGA-MENDOZA, R. M., LÓPEZ-MONROY, A. P., FRANCO-ARCEGA, A., MONTES-Y GÓMEZ, M. Emphasizing personal information for author profiling: new approaches for term selection and weighting. *Knowledge-Based Systems* 145 (2018), 169–181.

- [137] OUHBI, B., KAMOUNE, M., FRIKH, B., ZEMMOURI, E. M., BEHJA, H. A hybrid feature selection rule measure and its application to systematic review. In *Proceedings of the International Conference on Information Integration and Web-based Applications and Services* (2016), p. 106–114.
- [138] PARLAR, T., ÖZEL, S. A., SONG, F. A new feature selection method for sentiment analysis of Turkish reviews. In *Proceedings of the International Symposium on INnovations in Intelligent SysTems and Applications* (2016), p. 1–6.
- [139] PASHAEI, E., AYDIN, N. Binary black hole algorithm for feature selection and classification on biological data. *Applied Soft Computing Journal* 56 (2017), 94–106.
- [140] PATIL, L. H., ATIQUE, M. A novel feature selection based on information gain using WordNet. In *Proceedings of the Science and Information Conference* (2013).
- [141] PENNINGTON, J., SOCHER, R., MANNING, C. D. Glove: global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (2014), p. 1532–1543.
- [142] PEREIRA, R. B., PLASTINO, A., ZADROZNY, B., MERSCHMANN, L. H. C. Categorizing feature selection methods for multi-label classification. *Artificial Intelligence Review* 49, 1 (2018), 57–78.
- [143] PINHEIRO, R. H. W., CAVALCANTI, G. D. C., REN, T. I. Data-driven global-ranking local feature selection methods for text categorization. *Expert Systems with Applications* (2015).
- [144] PINTAS, J. T., CORREIA, L., BICHARRA GARCIA, A. C. Crowd-based feature selection for document retrieval in highly demanding decision-making scenarios. *Procedia Computer Science* 112 (2017), 822–832.
- [145] PRAMOKCHON, P., PIAMSA-NGA, P. A feature score for classifying class-imbalanced data. In *Proceedings of the International Computer Science and Engineering Conference* (2014).
- [146] QAZI, A., GOUDAR, R. H. An ontology-based term weighting technique for web document categorization. *Procedia Computer Science* 133 (2018), 75–81.
- [147] QIN, S., SONG, J., ZHANG, P., TAN, Y. Feature selection for text classification based on part of speech filter and synonym merge. In *Proceedings of the International Conference on Fuzzy Systems and Knowledge Discovery* (2016).
- [148] RAJAMOHANA, S. P., UMAMAHESWARI, K., KEERTHANA, S. V. An effective hybrid cuckoo search with harmony search for review spam detection. In *Proceedings of the IEEE International Conference on Advances in Electrical and Electronics, Information, Communication and Bio-Informatics* (2017).
- [149] RASOOL, A., TAO, R., KAMYAB, A. GAWA – a feature selection method for hybrid sentiment classification. *IEEE Access* 8 (2020), 191850–191861.

- [150] RASTOGI, S. Improving classification accuracy of automated text classifiers. In *Proceedings of the International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)* (2018), p. 1–7.
- [151] RAVI, K., RAVI, V. Sentiment classification of Hinglish text. In *Proceedings of the International Conference on Recent Advances in Information Technology* (2016).
- [152] REHMAN, A., JAVED, K., BABRI, H. A. Feature selection based on a normalized difference measure for text classification. *Information Processing and Management* 53, 2 (2017), 473–489.
- [153] REHMAN, A., JAVED, K., BABRI, H. A., ASIM, N. Selection of the most relevant terms based on a max-min ratio metric for text classification. *Expert Systems with Applications* 114 (2018), 78–96.
- [154] REHMAN, A., JAVED, K., BABRI, H. A., SAEED, M. Relative discrimination criterion - a novel feature ranking method for text data. *Expert Systems with Applications* (2015).
- [155] REN, J. S., WANG, W., WANG, J., LIAO, S. S. Exploring the contribution of unlabeled data in financial sentiment analysis. *arXiv preprint arXiv:1308.0658* (2013), 1149–1155.
- [156] RENNIE, J. The 20 newsgroups data set, 2019. Available at <http://qwone.com/~jason/20Newsgroups/>.
- [157] REZAEINIA, S. M., GHODSI, A., RAHMANI, R. Improving the accuracy of pre-trained word embeddings for sentiment analysis. *arXiv preprint arXiv:1711.08609* (2017).
- [158] ROUL, R. K., BHALLA, A., SRIVASTAVA, A. Commonality-rarity score computation. In *Proceedings of the Annual Meeting of the Forum on Information Retrieval Evaluation* (2016).
- [159] ROUL, R. K., GUGNANI, S., KALPESHBHAI, S. M. Clustering based feature selection using extreme learning machines for text classification. In *Proceedings of the IEEE International Conference Electronics, Energy, Environment, Communication, Computer, Control* (2016).
- [160] RUI, W., LIU, J., JIA, Y. Unsupervised feature selection for text classification via word embedding. In *Proceedings of the IEEE International Conference on Big Data Analysis* (2016), p. 1–5.
- [161] RUTA, D. Robust method of sparse feature selection for multi-label classification with naive Bayes. In *Proceedings of the Federated Conference on Computer Science and Information Systems* (2014), p. 375–380.
- [162] RZENIEWICZ, J., SZYMANSKI, J. S. Selecting features with SVM. In *Proceedings of the Iberoamerican Congress on Pattern Recognition* (2013).
- [163] SABBAH, T., SELAMAT, A., SELAMAT, M. H., IBRAHIM, R., FUJITA, H. Hybridized term-weighting method for dark web classification. *Neurocomputing* (2016).

- [164] SALGANIK, M. J., LEVY, K. E. Wiki surveys: Open and quantifiable social data collection. *PloS one* 10, 5 (2015), e0123483.
- [165] SAMMUT, C., WEBB, G. I., Eds. *Encyclopedia of machine learning*. Springer US, 2010.
- [166] SARHAN, A. M., HAMISSA, G. M., ELBEHIRY, H. E. Proposed document frequency technique for minimizing dataset in web crawler. In *Proceedings of the International Conference on Computer Engineering and Systems* (2016).
- [167] SHAH, F. P., PATEL, V. A review on feature selection and feature extraction for text classification. In *Proceedings of the IEEE International Conference on Wireless Communications, Signal Processing and Networking* (2016).
- [168] SHAHID, R., JAVED, S. T., ZAFAR, K. Feature selection based classification of sentiment analysis using biogeography optimization algorithm. In *Proceedings of the International Conference on Innovations in Electrical Engineering and Computational Technologies* (2017).
- [169] SHANG, C., LI, M., FENG, S., JIANG, Q., FAN, J. Feature selection via maximizing global information gain for text classification. *Knowledge-Based Systems* (2013).
- [170] SHANG, L., ZHOU, Z., LIU, X. Particle swarm optimization-based feature selection in sentiment classification. *Soft Computing* (2016).
- [171] SHEN, K., CHEN, X., KE, L., LU, Y., ZHANG, K. A blended feature selection method in text. In *Proceedings of the Conference on Cyberspace Technology* (2013), p. 573–576.
- [172] SHEYDAEI, N., SARAEE, M., SHAHGHOLIAN, A. A novel feature selection method for text classification using association rules and clustering. *Journal of Information Science* (2015).
- [173] SOMANTRI, O., KURNIA, D. A., SUDRAJAT, D., RAHANINGSIH, N., NURDIAWAN, O., PERDANA WANTI, L. A hybrid method based on particle swarm optimization for restaurant culinary food reviews. In *Proceedings of the International Conference on Informatics and Computing* (2019), p. 1–5.
- [174] SONG, J., ZHANG, P., QIN, S., GONG, J. A method of the feature selection in hierarchical text classification based on the category discrimination and position information. In *Proceedings of the International Conference on Industrial Informatics - Computing Technology, Intelligent Technology, Industrial Information Integration* (2016).
- [175] SONG, Q., NI, J., WANG, G. A fast clustering-based feature subset selection algorithm for high-dimensional data. *IEEE Transactions on Knowledge and Data Engineering* (2013).
- [176] SPOLAÔR, N., TSOUMAKAS, G. Evaluating feature selection methods for multi-label text classification. *BioASQ workshp* (2013).

- [177] STAMBAUGH, C., YANG, H., BREUER, F. Analytic feature selection for support vector machines. In *Proceedings of the Machine Learning and Data Mining in Pattern Recognition* (2013), p. 219–233.
- [178] STUDENT. The probable error of a mean. *Biometrika* (1908), 1–25.
- [179] SU, Z., XU, H., ZHANG, D., XU, Y. Chinese sentiment classification using a neural network tool - Word2vec. In *Proceedings of the International Conference on Multisensor Fusion and Information Integration for Intelligent Systems* (2014).
- [180] SUN, J., ZHANG, X., LIAO, D., CHANG, V. Efficient method for feature selection in text classification. In *Proceedings of International Conference on Engineering and Technology* (2017), vol. 2018-Janua, p. 1–6.
- [181] SUNDARARAJAN, K., PALANISAMY, A., VERSACI, M. Multi-rule based ensemble feature selection model for sarcasm type detection in Twitter. *Computational Intelligence and Neuroscience 2020* (2020), 2860479.
- [182] SUROWIECKI, J. *The wisdom of crowds*. Anchor, 2005.
- [183] TANG, B., HE, H. FSMJ: feature selection with maximum Jensen-Shannon divergence for text categorization. In *Proceedings of the World Congress on Intelligent Control and Automation* (2016), vol. 2016-Septe, p. 3143–3148.
- [184] TANG, B., HE, H., BAGGENSTOSS, P. M., KAY, S. A Bayesian classification approach using class-specific features for text categorization. *IEEE Transactions on Knowledge and Data Engineering* 28, 6 (2016), 1602–1606.
- [185] TANG, B., KAY, S., HE, H. Toward optimal feature selection in naive Bayes for text categorization. *IEEE Transactions on Knowledge and Data Engineering* (2016).
- [186] TANG, B., KAY, S., HE, H., BAGGENSTOSS, P. M. EEF: exponentially embedded families with class-specific features for classification. *IEEE Signal Processing Letters* (2016).
- [187] TANG, J., ALELYANI, S., LIU, H. Feature selection for classification: a review. *Data Classification: Algorithms and Applications* (2014).
- [188] TANG, X., DAI, Y., XIANG, Y. Feature selection based on feature interactions with application to text categorization. *Expert Systems with Applications* 120 (2019), 207–216.
- [189] TIAN, W., LI, J., LI, H. A method of feature selection based on Word2Vec in text categorization. In *Proceedings of the Chinese Control Conference* (2018), p. 9452–9455.
- [190] TOMMASEL, A. Integrating social network structure into online feature selection. In *Proceedings of the IJCAI International Joint Conference on Artificial Intelligence* (2016), vol. 2016-Janua, p. 4032–4033.
- [191] TRIPATHY, A., ANAND, A., RATH, S. K. Document-level sentiment classification using hybrid machine learning approach. *Knowledge and Information Systems* 53, 3 (2017), 805–831.

- [192] TRIVEDI, S. K., TRIPATHI, A. Sentiment analysis of Indian movie review with various feature selection techniques. In *Proceedings of the IEEE International Conference on Advances in Computer Applications* (2017).
- [193] TUTKAN, M., GANIZ, M. C., AKYOKUŞ, S. Helmholtz principle based supervised and unsupervised feature selection methods for text mining. *Information Processing and Management* (2016).
- [194] UYSAL, A. K. An improved global feature selection scheme for text classification. *Expert Systems with Applications* (2016).
- [195] UYSAL, A. K., GUNAL, S. A novel probabilistic feature selection method for text classification. *Knowledge-Based Systems* 36 (2012), 226–235.
- [196] VANI, K., GUPTA, D. Text plagiarism classification using syntax based linguistic features. *Expert Systems with Applications* 88 (2017), 448–464.
- [197] VIJAYARANI, S., ILAMATHI, M. J., NITHYA, M. Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks* 5, 1 (2015), 7–16.
- [198] VYCHEGZHANIN, S. V., RAZOVA, E. V., KOTELNIKOV, E. V. What number of features is optimal: a new method based on approximation function for stance detection task. In *Proceedings of the International Conference on Information Communication and Management* (2019), ICICM 2019, p. 43–47.
- [199] W3TECHS. Historical trends in the usage of content languages for websites, September 2019, 2019.
- [200] WANG, D., ZHANG, H., LIU, R., LIU, X., WANG, J. Unsupervised feature selection through Gram-Schmidt orthogonalization - a word co-occurrence perspective. *Neurocomputing* (2016).
- [201] WANG, D., ZHANG, H., LIU, R., LV, W., WANG, D. T-test feature selection approach based on term frequency for text categorization. *Pattern Recognition Letters* (2014).
- [202] WANG, H., HONG, M. Supervised Hebb rule based feature selection for text classification. *Information Processing & Management* 56, 1 (2019), 167–191.
- [203] WANG, J., WU, L., KONG, J., LI, Y., ZHANG, B. Maximum weight and minimum redundancy: a novel framework for feature subset selection. *Pattern Recognition* (2013).
- [204] WANG, Q., LIU, L., JIANG, J., JIANG, M., LU, Y., PEI, Z. Feature selection method based on multiple centrifuge models. *Cluster Computing* 20, 2 (2017), 1425–1435.
- [205] WANG, Y., LIU, Y., FENG, L., ZHU, X. Novel feature selection method based on harmony search for email classification. *Knowledge-Based Systems* (2014).
- [206] WANG, Y., LIU, Y., ZHU, X. Two-step based hybrid feature selection method for spam filtering. *Journal of Intelligent & Fuzzy Systems* 27 (2014), 2785–2796.



- [207] WANG, Y., WANG, J., LIAO, H., CHEN, H. An efficient semi-supervised representatives feature selection algorithm based on information theory. *Pattern Recognition* (2017).
- [208] WEBKB. The 4 universities data set, 2019. Available at <http://www.cs.cmu.edu/afs/cs/project/theo-20/www/data/>.
- [209] WOLD, S., ESBENSEN, K., GELADI, P. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems 2*, 1-3 (1987), 37–52.
- [210] WU, G., WANG, L., ZHAO, N., LIN, H. Improved expected cross entropy method for text feature selection. In *Proceedings of the International Conference on Computer Science and Mechanical Automation* (2016).
- [211] WU, G., XU, J. Optimized approach of feature selection based on information gain. In *Proceedings of the International Conference on Computer Science and Mechanical Automation* (2016).
- [212] WU, L., WANG, Y., ZHANG, S., ZHANG, Y. Fusing gini index and term frequency for text feature selection. In *Proceedings of the IEEE International Conference on Multimedia Big Data* (2017).
- [213] XIAOMING, D., TANG, Y. Improved mutual information method for text feature selection. In *Proceedings of the International Conference on Computer Science & Education* (2013).
- [214] XU, H., XU, L. Multi-label feature selection algorithm based on label pairwise ranking comparison transformation. In *Proceedings of the International Joint Conference on Neural Networks* (2017).
- [215] XU, J., JIANG, H. An improved information gain feature selection algorithm for SVM text classifier. In *Proceedings of the International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery* (2015).
- [216] XU, Z., KING, I., LYU, M., JIN, R. Discriminative semi-supervised feature selection via manifold regularization. *IEEE Transactions on Neural Networks 21* (2010), 1033–1047.
- [217] YANG, J., LIU, Z., QU, Z., WANG, J. Feature selection method based on crossed centroid for text categorization. In *Proceedings of the IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing* (2014).
- [218] YANG, J., LU, Y., LIU, Z. An improved strategy of the feature selection algorithm for the text categorization. In *Proceedings of the IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing* (2019), p. 3–7.
- [219] YANG, J., WANG, J., LIU, Z., QU, Z. A term weighting scheme based on the measure of relevance and distinction for text categorization. In *Proceedings of the IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing* (2015).

- [220] YANG, Y., PEDERSEN, J. O. A comparative study on feature selection in text categorization. In *Icml* (1997), vol. 97, p. 412–420.
- [221] YANG, Z.-T., ZHENG, J. Research on Chinese text classification based on Word2vec. In *Proceedings of the IEEE International Conference on Computer and Communications Research* (2016).
- [222] YIGIT, F., BAYKAN, O. K. A new feature selection method for text categorization based on information gain and particle swarm optimization. In *Proceedings of IEEE International Conference on Cloud Computing and Intelligence Systems* (2014).
- [223] YOUSEFPOUR, A., IBRAHIM, R., HAMED, H. N. A. Ordinal-based and frequency-based integration of feature selection methods for sentiment analysis. *Expert Systems with Applications* 75 (2017), 80–93.
- [224] ZAINUDDIN, N., SELAMAT, A., IBRAHIM, R. Hybrid sentiment classification on Twitter aspect-based sentiment analysis. *Applied Intelligence* 48, 5 (2018), 1218–1232.
- [225] ZHANG, H., REN, Y. G., YANG, X. Research on text feature selection algorithm based on information gain and feature relation tree. In *Proceedings of the Web Information System and Application Conference* (2013), p. 446–449.
- [226] ZHANG, J., HU, X., LI, P., HE, W., ZHANG, Y., LI, H. A hybrid feature selection approach by correlation-based filters and SVM-RFE. In *Proceedings of the International Conference on Pattern Recognition* (2014), p. 3684–3689.
- [227] ZHANG, M.-L., ZHOU, Z.-H. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering* 26, 8 (2014), 1819–1837.
- [228] ZHANG, Z., KE, T., DENG, N., TAN, J. Biased p-norm support vector machine for PU learning. *Neurocomputing* 136 (2014), 256–261.
- [229] ZHEN, Z., WANG, H., XING, Y., HAN, L. Text feature selection approach by means of class difference. In *Proceedings of the International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery* (2016).
- [230] ZHOU, H., GUO, J., WANG, Y., ZHAO, M. A feature selection approach based on interclass and intraclass relative contributions of terms. *Computational Intelligence and Neuroscience* 2016 (2016).
- [231] ZHOU, H., HAN, S., LIU, Y. A novel feature selection approach based on document frequency of segmented term frequency. *IEEE Access* 6 (2018), 53811–53821.
- [232] ZHOU, X., HU, Y., GUO, L. Text categorization based on clustering feature selection. *Procedia Computer Science* 31 (2014), 398–405.
- [233] ZHU, L., WANG, G., ZOU, X. Improved information gain feature selection method for Chinese text classification based on word embedding. In *Proceedings of the International Conference on Software and Computer Applications* (2017).

- 
- [234] ZHUANG, Y., WANG, H., XIAO, J., WU, F., YANG, Y., LU, W., ZHANG, Z. Bag-of-discriminative-words (BoDW) representation via topic modeling. *IEEE Transactions on Knowledge and Data Engineering* 29, 5 (2017), 977–990.
  - [235] ZONG, W., WU, F., CHU, L. K., SCULLI, D. A discriminative and semantic feature selection method for text categorization. *International Journal of Production Economics* 165 (2015), 215–222.
  - [236] ZUO, Z., LI, J., ANDERSON, P., YANG, L., NAIK, N. Grooming detection using fuzzy-rough feature selection and text classification. In *Proceedings of the IEEE International Conference on Fuzzy Systems* (2018), p. 1–8.

## Appendix A. List of Acronyms

		<b>DT</b>	Decision Tree
<b>ACC</b>	Accuracy Measure	<b>EEFS</b>	Ensemble Embedded Feature Selection
<b>ACC2</b>	Balanced Accuracy Measure	<b>FRFS</b>	Fuzzy Rough Feature Selection
<b>ALOFT</b>	At Least One FeaTure	<b>FS</b>	Feature Selection
<b>ANOVA</b>	Analysis of Variance	<b>GAWA</b>	Genetic Algorithm and Wrapper Approaches
<b>BACA</b>	Bit-priori Association Classification Algorithm	<b>GFSS</b>	Global Filter-based Feature Selection Scheme
<b>BBHA</b>	Binary Black Hole Algorithm	<b>GI</b>	Gini Index
<b>BFSM</b>	Blended Feature Selection Method	<b>GPSO</b>	Geometric Particle Swarm Optimization
<b>BGSA</b>	Binary Gravitational Search Algorithm	<b>HAN</b>	Hierarchical Attention Network
<b>BMI</b>	Balanced Mutual Information	<b>HRFS</b>	Hebb Rule Based Feature Selection
<b>BoDW</b>	Bag of Discriminative Words	<b>IDF</b>	Inverse Document Frequency
<b>BoW</b>	Bag of Words	<b>IG</b>	Information Gain
<b>BPSO</b>	Binary Particle Swarm Optimization	<b>IPSO</b>	Improved Particle Swarm Optimization
<b>CAS</b>	Correlative Association Score	<b>ISCA</b>	Improved Sine Cosine Algorithm
<b>CDM</b>	Class Discriminating Measure	<b>KNN</b>	$k$ -Nearest Neighbors
<b>CHI</b>	Chi-square	<b>LDA</b>	Latent Dirichlet Allocation
<b>CMFS</b>	Comprehensively Measure Feature Selection	<b>LSAN</b>	Latent Selection Augmented Naive Bayes
<b>CNN</b>	Convolutional Neural Network	<b>MBF</b>	Markov Blanket Filter
<b>CrowdFS</b>	Crowd-based Feature Selection	<b>MFS</b>	Meta Feature Selection
<b>CSO</b>	Cat Swarm Optimization	<b>MFSLFD</b>	Memetic Feature Selection based on Label Frequency Difference
<b>DBN</b>	Deep Belief Network	<b>MI</b>	Mutual Information
<b>DF</b>	Document Frequency	<b>MMI</b>	Multivariate Mutual Information
<b>DFS</b>	Discriminative Features Selection	<b>MMR</b>	Max-Min Ratio
<b>DFS*</b>	Distinguishing Feature Selector	<b>MOANOF</b>	Multi-Objective Automated Negotia-
<b>DGBFS</b>	Diversified Greedy Backward-Forward Search		
<b>DPP</b>	Discriminative Personal Purity		

	tion based Online Feature Selection	<b>SAIG</b>	Sparsity Adjusted Information Gain
<b>MORDC</b>	Multi-Objective Relative Discriminative Criterion	<b>SBATFS</b>	Spark BAT Feature Selection
<b>MRDC</b>	Multivariate Relative Discrimination Criterion	<b>SIGCHI</b>	Square of Information Gain and Chi-square
<b>NB</b>	Naive Bayes	<b>SLR</b>	Systematic Literature Review
<b>NDM</b>	Normalized Difference Measure	<b>SMOTE</b>	Synthetic Minority Oversampling Technique
<b>OR</b>	Odds Ratio	<b>SVM</b>	Support Vector Machines
<b>OS-FS</b>	Optimized Swarm Search-based Feature Selection	<b>SVM-RFE</b>	Support Vector Machine-Recursive Feature Elimination
<b>PCT</b>	Pairwise Comparison Transformation	<b>SWA</b>	Small World Algorithm
<b>POS</b>	Part of Speech	<b><i>t</i>-Test</b>	Student's <i>t</i> -Test
<b>POSFilter</b>	Part of Speech Filter	<b>TF</b>	Term Frequency
<b>PSO</b>	Particle Swarm Optimization	<b>TF-IDF</b>	Term Frequency-Inverse Document Frequency
<b>RCV1</b>	Reuters Corpus Volume I	<b>WFSAIG</b>	Wrapper Feature Selection Algorithm based on Iterated Greedy
<b>RDC</b>	Relative Discrimination Criterion	<b>WI-OMFS</b>	Wolf Intelligence Based Optimization of Multi-Dimensional Feature Selection Approach
<b>RF</b>	Random Forest		
<b>RFE</b>	Recursive Feature Elimination		
<b>RP-GSO</b>	Random Projection and Gram-Schmidt Orthogonalization		