Text Classification Using Word Embedding in Rule-Based Methodologies: A Systematic Mapping

Asmaa M. Aubaid [1], Alok Mishra [2]

[1]*Department of Modeling & Design of Engineering Systems, Atilim University, Ankara, Turkey*
[2] *Department of Software Engineering, Atilim University, Ankara, Turkey*

Abstract **– With the advancing growth of the World Wide Web (WWW) and the expanding availability of electronic text documents, the automatic assignment of text classification (ATC) has become more important in sorting out information and knowledge. One of the most crucial tasks that should be carried out is document representation using word embedding and Rule-Based methodologies. As a result, this, along with their modeling methods, has become an essential step to improve neural language processing for text classification. In this paper, a systematic mapping study is a way to survey all the primary studies on word embedding to rule-based and machine learning of automatic text classification. The search procedure identifies 20 articles as relevant to answer our research questions. This study maps what is currently known about word embedding in rule-based text classification (TC). The result shows that the research is concentrated on some main areas, mainly in social sciences, shopping products classification, digital libraries, and spam filtering. The present paper contributes to the available literature by summarizing all research in the field of TC and it can be beneficial to other researchers and specialists in order to sort information.**

Keywords **– Systematic Mapping, Word Embedding, Rule-Based, Text Classification.**

DOI: 10.18421/TEM74-31
https://dx.doi.org/10.18421/TEM74-31

Corresponding author: Alok Mishra,
Department of Software Engineering, Atilim University, Ankara, Turkey
Email: alok.mishra@atilim.edu.tr

Received: 12 October 2018.
Accepted: 14 November 2018.
Published: 26 November 2018.

1. Abbreviations **–** WWW: World Wide Web; TC: Text Categorization; ML: Machine Learning; NLP: Natural Language Process; IR: Information Retriever; ARC-BC: Association Rule-based Classifier By Categories; ARTC: Association Rule-based Text Classifier Algorithm; AIRTC: Automatic Induction of Rule Based Text Categorization; SVM : Support Vector Machines Model; LSTM : Long Short Term Memory; SLR : Systematic Literature Review; ARQs: Addressing of Questions.

2. Introduction

The extensive volume of data and information accessible in digital form and the need to sort it has progressively intensified interest in automatic text categorization (ATC). Systematic mapping study is a type of auxiliary study intending to conduct an exhaustive review of a specific research topic, to identify the gaps, and to gather proof with the specific end-goal of directing future studies [1]. Also, secondary studies are drawing increasing attention from academia since, given the growing quantity of reports in any field, it is constantly growing and, there is a need to provide an overview of the current scientific sources, so as to set up a strong basis for further researches around any subject [2, 3]. A common way to do this is with word embedding, mainly with the help of rule-based and machine-learning techniques. Word embedding is a form of representation that enables words with similar meanings to also have a similar representation. This allows machines to develop a better understanding of words (ideas). Word embedding is regarded as one of the main challenges for Natural Language Process (NLP) & Information Retriever (IR) communities. Because of the easy interpretability of the standard rules, rule-based classification systems have been widely used in real world applications. The term 'rule-based' classification can be used to refer to any categorization schemes using IF-THEN rules to predict cretin classes. The upsides of this approach are to see and understand data direction in an essential and direct framework compared to machine learning. In any case, machine learning

represents a branch of synthetic intelligence and incorporates two phases: "training" and "testing" (classification). "Training" means to develop a classifier utilizing tools of classification. (e.g., vector space model method). "Testing" means to ensure that documents are properly categorized by the classifiers.

3. Organization of the Paper

This study contains ten sections to cover:

Section: 1 contains the abbreviations of terms of this systematic mapping**.**

Section: 2 is an introduction to automatic text categorization approach, rule-based, ML and related works.

Section: 3 includes the outline of the article.

Section:4 contains a brief review of word-embedding techniques.

Section: 5 introduces rule-based inadvertent forms, such as Association Rule-based Classifier by Categories (ARC-BC), Automatic Induction of Rule-Based Text Categorization (AIRTC) and Automatic Induction of Rule-Based Text Categorization (AIRTC).

Section:6 provides a definition of Machine Learning Models (ML) approach with Support Vector Machines Model (SVM), Convolution Neural Network (CNN) and Long Short Term Memory (LSTM).

Section:7 contains the systematic mapping study and its comparison with the literature survey.

Section:8 splits the procedure into three fundamental stages: research orders, data accumulation and results.

Section: 9 offers the motivation behind this study with sub-sections containing the goal, screening of papers for inclusion and exclusion, and questions.

Section: 10 addresses the questions, discussion, conclusion, and future work.

4. Word Embedding

Word embedding is defined as content representation in a way that the words which have similar meanings also share the same representation. This is the way to deal with representing words and archives, and may be seen as one of the key steps ahead of deep learning when testing NLP problems. One of the benefits of utilizing dense and low-dimensional vectors is computational capacity since large portions of neural network tool kits cannot handle high-dimensional, dispersed vectors [4]. Word embedding is, in truth, a class of methods, where singular words are represented to real-valued vectors in a predefined vector space. Each word is mapped to one vector and the vector values are discovered in a way that takes after a neural network, and from this time forward the technique is generally within the field of profound learning. Key to the approach is using a densely dispersed representation for each word. a real-valued vector, frequently tens or more measurements are utilized to represent each word. This is divided into thousands or much larger numbers of dimensions required for inadequate word depictions, for instance, a one-hot encoding [5].

5. Rule - Based

Rule-based systems (often called "Generation Systems" or "Expert Systems") belong to artificial intelligence. A rule-based system utilizes rules as the learning representation for the information coded into the system [6,7, 8, 9]. The implications of rule-based systems depend completely on expert systems, which copy the reasoning of human experts in explaining an information-intensive issue. Instead of learning in a declarative, static and a way as a course action of things which are valid, rule-based systems represent knowledge in terms of a set of rules that determines what to do or what to conclude in various situations.

5.1. Association Rule-based Classifier by Categories (ARC-BC)

This model is created to form an already known text classifier. The ARC-BC algorithm considers each set of text documents; they have a place with one classification as a different content accumulation to produce association rules. On the off chance that the document has a place with more than one class, this document will be available in each set related with classifications that the document falls into. However, the algorithms may be unable to classify a single-class document that has a few terms of document mutually connected with another class [10].

5.2. Association Rule-Based Text Classifier Algorithm (ARTC)

This section presents another rule-based text classifier algorithm to enhance the forecast accuracy of Association Rule-based Classifier by Categories (ARC-BC) algorithm. Unlike the previous algorithms, the proposed association rule generation

algorithm constructs two kinds of successive item sets. The first frequent item set, i.e. L_k, also contains all features that overlap with other categories. The second frequent item set, i.e.OL_k contains all features that overlap with other categories. Further, this paper additionally proposes another operation for the second regular item sets. The exploratory results show acceptable performance by the proposed classifier [11].

5.3. Automatic Induction of Rule -Based Text Categorization (AIRTC)

In the literature, the common approach to classification depends on the machine learning systems: a general inductive process automatically assembles a classifier by learning, from a set of pre-classified documents, the qualities of the classifications. This section depicts a novel strategy for the automatic induction of rule-based text classifiers. This method supports a hypothesis language of the form "if T_1, … or T_n occurs in document d, and none of T_{1+n},... T_{n+m},occurs in d, then classify d under category c," where each T_i is a conjunction of terms. This rule is about the primary ways to deal with text classification within the machine-learning model. Issues relating to three distinct topics, to be specific, document representation, classifier development, and classifier assessment - are discussed - in detail [12].

6. Machine Learning Models (ML)

The definition of Machine Learning (ML) is "The ability of a machine to improve its performance based on previous results". Therefore, machine learning document classification is “the ability of a machine to improve its document classification performance based on previous results of document classification”. In machine learning, the goal of classification is to group the items that have similar feature values. In the following sections, we will explore some of the most popular models of ML.

6.1. Support Vector Machines Model (SVM)

Support Vector Machine (SVM) is a regulated machine learning algorithm which can be utilized for both classification and regression challenges. However, it is for the most part utilized as a part of classification issues. In this algorithm, we plot every datum item as a point in n-dimensional space (where n the number of terms) with the value of each term being the value of a particular coordinate. At this point, classification is done by finding the hyper-plane that separates the two classes entirely well (see Figure 1.) in which Support Vector Machine approach is illustrated.

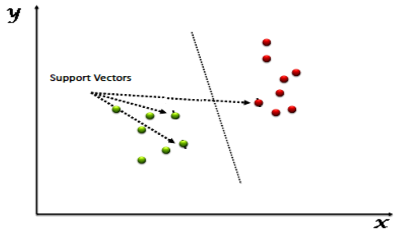


Figure 1. Support Vector Machine approach (SVM)

6.2. Convolution Neural Network (CNN)

We employ the basic Convolution Neural Network as proposed in [13], which has performed well in different sentence classification assignments. The network’s input is a sentence matrix X shaped by a linking k-dimensional word embedding. At this point, a convolution filter ($W \in R^{h\times k}\ i$) is implemented to every possible sequence of length h to get a feature map:

$$c_i = \tanh(W.X + b)\text{……………….……..(1)}$$

Followed by a max-over-time pooling operation to obtain the element with the highest quality:

$$\hat{c} = maxc\ \text{……………………………….(2)}$$

The pooled features of various filters are, then, linked and turned into completely associated soft-max layers to perform the classification. The network utilizes various filters with various sequence sizes also covering diverse sizes of windows in the sentence.

All the hyper parameters of the network are utilized in the same way as the original paper [14], with a stochastic dropout [15], p = 0.5 on the penultimate layer, and 100 filters for each filter region with a width of 2, 3 and 4. Enhancement is performed with Ad delta on mini-batches of size 50 [16].

6.3. Long Short Term Memory (LSTM)

The long short-term memory (LSTM) block or network is a basic recurrent neural network which can be utilized as a building part or block of hidden layers to obtain greater repetitive neural networks. The LSTM block itself is a recurrent network since it contains intermittent associations in an ordinary recurrent neural network. In the same way, does LSTM systems are repetitive neural systems where intermittent units comprise a memory cell c and three entryways I, o and f [17].Given an arrangement of information embedding x, LSTM yields a grouping of states h given by the accompanying conditions:

$$\begin{pmatrix} i_t \\ f_t \\ o_t \\ \tilde{c}_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ tanh \end{pmatrix} w. \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix} \text{.....(3)}$$

$$c_t = f_t \odot c_{t-1} + \mathrm{i_t} \odot \tilde{\mathrm{c}}_\mathrm{t} \text{ ……..…….…….(4)}$$

$$h_t = o_t \odot \tanh(c_t) \text{……(5)}$$

Where

$W \in R^{4k \times 2k}$, $c_t^{\sim}$ is a candidate state for the memory cell and _ ⊙ is an element-wise vector multiplication.

The conveyance of labels for the whole sentence is processed by a completely associated soft-max layer on top of the final hidden state in the wake of applying a stochastic dropout with p = 0.25. It is used in 150 dimensions for the size of (h), Ad grad for advancement and mini-batch size of 100 [18].

7. Systematic Mapping Study

In order to systematically manage word embedding in rule-based and ML of ATC, it is necessary to have a clear and thorough understanding of the state of the ATC. Different methods and tools have been used, proposed, and developed for ATC, but it is not clear how these methods and tools map ATC activities. This paper reports the results of a systematic mapping study, broadly examining the concept in (ATC) and its management. Systematic mapping study is an of optional approach to extensive review specific research topic, identify gaps, and gather proof so as to direct future research for the good idea [19]. Studies like this are drawing more and more attention from the scientific community because, when the quantity of reports in a field is always developing, it becomes plainly vital to provide a review of the current logical sources to structure them and gather satisfactory support for additional investigations of the subject [20, 21].

7.1. Systematic Mapping Study and Systematic Literature Review (SLR)

Systematic mapping is an activity frequently performed in medical research; it gives a structure to research reports and results that have been distributed by means of categorizing them. It offers a visual synopsis, which is a map, of its results. Systematic mapping requires less exertion while offering a more detail review. Proof - based drugs have been a recent case of using secondary studies, particularly in the systematic literature review (SLR) approach [19]. This entails going in depth within the existing studies while keeping in mind the end goal to survey their significance and identify their specific methodology [19]. As indicated by Kitchenham et al. [19], this approach is quite common for such fields as criminology, social studies, finance, nursing and so on, where only an accumulation of every current proof about some predefined subject is not sufficient in light of the fact that there is a need for thorough technique which maintains a strategic distance from any inclination and error in source selection and review. Keele [20] speaks of this approach as "a means of evaluating and interpreting all available research relevant to a particular research question, topic area, or phenomenon of interest". One of such techniques is systematic mapping study, which "gives a review of a research area, and identifies the quantity and type of research and results available within it" [21].

7.2. The difference between Systematic Mapping Study and Systematic Literature Review

There are contrasts between a mapping study and SLR. Keele [20] summarizes them as follows:

- Research in queries of mapping studies are typically more extensive and often numerous.
- The indexed lists for mapping studies are probably going to restore a major number of studies. In any case, it is not as risky as it is for SLR because the aim is broad coverage rather than narrow focus.

8. Stages of a Mapping Study

As can be found in Fig. 2., the procedure is clearly divided into three fundamental stages:

1.Research orders
2.Data accumulation
3.Results.

This is in accordance with the practices of systematic reviews [20], which determine planning, conducting and reporting stages.

These stages are named differently in contrast to what is characterized for systematic reviews; however, the general idea and aim for each stage is followed. First, the protocol and the research questions are determined. This is an essential stage since research objectives are according to the answers to these questions. The second stage contains the execution of the mapping study, in which we look for essential investigations. A set of inclusion and exclusion criteria is arranged and utilized as a part of the selection procedure in order to find studies that may contain relevant results as

per the objectives of the research. In the third stage, a classification scheme is produced. This is done with two aims in mind, to organize the subject in accordance with the research questions, and to consider different research types as characterized in [18]. In the end, the results are gained best on the extensive investigation from a mapping study with separate stages as in Figure 2.

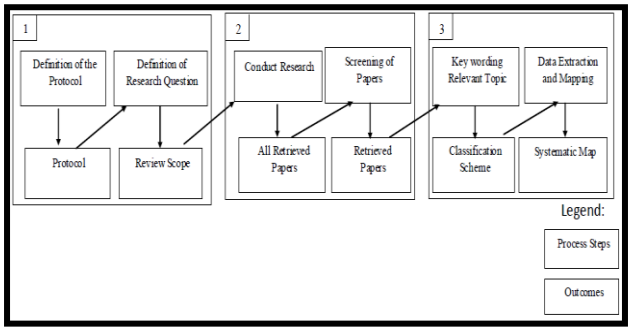


Figure 2. The systematic mapping process (adapted from Petersen et al. [18]).

9. Motivation of the Systematic Mapping Study

9.1. Goal of the Study

The objective of this investigation, depicted by utilizing the Goal-Question-Metric approach (Basili,1992), is: to analyze the main studies on ATC to get an in-depth view regarding ATC, word embedding and learning machine algorithms from the perspective of scientists and experts with regards to content of text classification development. Before going into the mapping study process, ten research questions are developed. These questions determine the objective of the present work, which ultimately attempts to identify the natures of different articles for these questions.

9.2. Screening of Papers for Inclusion and Exclusion

The inclusion and exclusion criteria are used to exclude contributions that are not applicable to answer the questions of this research. The authors believe that it is necessary to eliminate papers, which have only specified our main focus in initial sentences in theory. This was required since it is a focal idea in the area and consequently, is every now and again utilized as abstracts without papers, not quite addressing it any further. We prototyped this technique and did not find any misclassifications as a result [21].

9.3. View of the Questions

We explore answers to these ten questions in the 20 articles, and then classify these papers accordingly. Table 1. shows the research questions and the answers related to them.

Table 1. The research questions with their answers.

Research Questions	Motivation of Questions
RQ1: What are the types of studies?	For every study, it is needed to carry out some experiments or simulations to obtain results, implying that there are both empirical and experimental studies.
RQ2: What are the domain areas of these papers?	This question represents the fields of interest such as IT, computer engineering, libraries, and agencies.
RQ3: What are the purposes of the studies?	Some studies or researches intend to classify text documents in the dataset according to their categories, and spam message.
RQ4: Which countries do the studies originate from?	The country of the authors for example, Belgium, Mexico, etc**.**
RQ5: What are the methods used in the papers?	Researchers apply many methods depending on the word embedding and learning machine in order to obtain satisfactory results and choose the best among them, such as Word2Vec, GloVe , Naïve Bayes,etc**.**
RQ6: What are the criteria for validity?	To determine the quality of performance in a study and select the best one.
RQ7: Which journals have published these studies?	Whether science journals, or other forms of publication forums.
RQ8: What are the acronyms used in the study manuscripts?	Different abbreviations such as NLP (Natural Language Processing), ML (Machine Learning),etc.
RQ9: Which years and journals have the highest number of articles related to the topic of our study?	The ACM journal has the highest number of publications in 2017.
RQ10: Which models of text classification have been used in the studies?	There are a few models as a part of investigations of this paper, for example Global Vectors and ARTC. It can classify studies according to their models.

- **RQ1: What are the types of studies?**

For each study, it is needed that we implement some experiments or simulations to obtain acceptable results, thus classifying the study into empirical and experimental types. Table 2. shows the two types of studies in the articles

Table 2. The two types of studies in the articles

Articles	Answers
Std1, Std2, Std3, Std4, Std6, Std7, Std8, Std9, Std10, Std11, Std12, Std13, Std16, Std17, Std19, Std20.	Experimental study.
Std5, Std14, Std15, Std18.	Empirical study.

- **RQ2: What are the domain areas?**

This question represents the field in which the studies are most interested, such as IT, computer engineering, libraries, agents, etc. Therefore, Table 3. explains the domain of the studies.

Table 3. The domain of the studies

Articles	Answers
Std1	Evaluation of genetic test datasets.
Std2	Biomedical literature indexing.
Std3	Retrieval of electronic documents from digital libraries
Std4	Classifying models in Chinese web pages
Std5	Patent classification
Std6	Disambiguation of texts in the Marathi Language
Std7	Improving the performance of text categorization of benchmark data collections
Std8	Topic segmentation in Arabic and English
Std9	Detecting phishing attacks in Internet banking.
Std10	Early detection of gradual concept drifts by text categorization.
Std11	Information retrieval and text categorization in Chinese and English document collections
Std12	Sentiment level classification in English documents
Std13	Using Topic2Vec in the same semantic vector space with words.
Std14	Improving the accuracy of a near-state-of-the-art supervised NLP system
Std15	Improving the efficiency of ML
Std16	Using a classification of natural disasters news reports in Spanish newspaper articles.
Std17	Intrusion Detection Systems.
Std18	Reviewing datasets of movies on the Amazon webpage.
Std19	Implementing labeled or unlabeled data
Std20	Data mining in hospitals and healthcare centers for data mining.

- **RQ3: What are the purposes of the studies?**

Some researchers classify text documents in datasets according to their categories and remove spam message. As such, Table 4. explains the purpose of the studies.

Table 4. Purpose of studies

Articles	Answers
Std1	Learning lists of meta-rules to generalize the choice of the best classifier for text datasets.
Std2	An original neural language model is proposed, that is topic-based skip-gram, to learn topic-based word embedding in the field of biomedicine by indexing with CNNs.
Std3	This work addresses the various methods for document representation in the form of a fully-inverted index as the basis for operations on string vectors
Std4	An ensemble classifier is introduced combining classification rules and the statistical language model.
Std5	This paper presents a patent categorization method in accordance to word embedding and long short-term memory networks in order to categorize patents down to the subgroup IPC level.
Std6	A Rule-Based method is proposed to carry out Word Sense Disambiguation in Marathi Language.
Std7	An n vector space model (VSM) is suggested so that the documents can be recognized and classified by a computer or a classifier.
Std8	LSA, Word2Vec and GloVe methods are used in the field of topic segmentation for Arabic and English.
Std9	A new application of rule-based method identifies phishing attacks on Internet banking websites.
Std10	Here, a TRIO algorithm for online detection of signal drifts is suggested.
Std11	This paper comparatively examines TF_IDF, LSI and multi-word for text representation.
Std12	A new framework is used to capture sentiment information of various types.
Std13	Topic2Vec approach is used to learn topic representations in the same semantic vector space with words, as an option for probability distribution.
Std14	In this study, Brown clusters, Colbert, Weston embedding, and HLBL embedding of words are investigated on both NER and chunking.
Std15	This paper reviews the possibility of upgrading the conventional "bag_of_words" model to shed light on the structural features of text documents and to consider them in the process of categorization with the help of machine learning theory methods.
Std16	A new semi-supervised method for text categorization is offered.
Std17	In this attempt, network-based IDS are examined by combining two data mining algorithms, C4.5 Decision Tree and SVM.
Std18	This paper studies CNN on text categorization for use in 1D structure (namely, word order) of text data so as to make accurate predictions possible.
Std19	An effective and efficient use of LSTM is displayed in this work within supervised and semi-supervised settings.
Std20	This is a comparison of techniques and experiments in the sub domain of radiology text classification.

- **RQ4: Which countries do the studies originate from?**

Some countries are interested in Automatic Text Classification (ATC) with different techniques, such as Belgium, Mexico…etc. Table 5. gives an example of countries supporting ATC studies.

Table 5. Countries of studies

Articles	Countries	Articles	Countries
Std1	Belgium, Mexico.	**Std11**	China.
Std2	USA.	**Std12**	USA, Singapore, China.
Std3	India.	**Std13**	China.
Std4	China.	**Std14**	Canada.
Std5	Brazil.	**Std15**	Moscow, Russia
Std6	India.	**Std16**	Mexico , Spain.
Std7	China, Singapore.	**Std17**	India.
Std8	Tunis.	**Std18**	USA.
Std9	Iran.	**Std19**	USA , China.
Std10	Italy.	**Std20**	USA.

- **RQ5: What are the methods used in the papers?**

There some methods used in different contributions such as Word2Vec, GloVe, RIPPER, C4.5, Naïve Bayes, SVM, etc. Table 6. shows many types of models used in studies.

Table 6. The models used in the articles

Articles	Answers
Std1	Evolutionary Learning of Meta-Rules (ELMR).
Std2	Skip-gram, Convolution Neural Networks (CNNs) and word embedding.
Std3	Support vector machines (SVM), Decision Tree and Neural Network Classifiers.
Std4	Strong Covering Algorithm (SCA), Statistical Language Model.
Std5	Long Short Term Memory (LSTM) and Word2Vec.
Std6	Rule -Based method.
Std7	Term Frequency Support vector machines (SVM), Inverse Document Frequency (TF-IDF) and KNN.
Std8	Latent Semantic Analysis (LSA), Global Vectors (GloVe) and Word2Vec.
Std9	Rule- based and Support vector machines (SVM) .
Std10	TRIO algorithm.
Std11	Support vector machines (SVM) , LSI and TF-DF.
Std12	GloVe Algorithm and CBOW
Std13	Neural Probabilistic Language Model (NPLM) , Bag-of-Words (BOW) and Continuous Bag-of-Words (CBOW).
Std14	Hierarchical log-bilinear (HLBL) model.
Std15	SVM , NB, LR , KNN , C4.5 Algorithm and AdaBoost
Std16	Support Vector Machines (SVM) and Naïve Bayes (NB) techniques.
Std17	Support Vector Machine (SVM) and C4.5 decision tree methods.
Std18	Support Vector Machine (SVM).
Std19	Long Short-Term Memory (LSTM).
Std20	Rule-Based Method.

- **RQ6: What are the criteria for validity?**

Recognizing the performance of a research relies upon detecting the validity standards, and the best study is considered as having the highest validity as a result. Table 7. explains the validity criteria satisfied in the different studies

Table 7. Validity criteria satisfied in the studies.

Articles	Answers
Std1	The experiments in this study show encouraging results, because it uses the Evolutionary Learning of Meta-Rules (ELMR) method for text classification.
Std2	Topic-based semantic word embeddings with multimodal CNNs outperform state-of-the-art word representations in text classification.
Std3	Better than the traditional approach to represent the document by minimizing document pre-processing time and feature dimensionality. Also, potential ease is offered for tracing why each document is classified under a certain category.
Std4	The experimental result shows that the classifying models in Chinese web page classifications are better than traditional rule-based and statistical classifying models.
Std5	The classification method achieves 63% accuracy at the subgroup level.
Std6	The correct sense of the given text is detected from the predefined conceivable senses utilizing word principles and sentence rules.
Std7	Supervised term weighting method, tf-idf, has a superior advantage over other term-weighting techniques. .
Std8	Word2Vec presents an ideal word vector portrayal.
Std9	The proposed feature sets along with others can detect phishing pages in Internet banking with accuracy of 99.14% true positive and only 0.86% false negative alarm.
Std10	The validity is confirmed by comparison with a case in the literature concerning increasing or decreasing signals.
Std11	Experimental results reveal that in TC, LSI shows better execution over different techniques in the two document collections.
Std12	Using earlier sentiment knowledge into the embedding procedure can lead to learning better representations for sentiment analysis.
Std13	Topic2Vec yields unexpected and satisfactory results.
Std14	Each of the three word-representations can enhance the accuracy of the baselines.
Std15	The efficiency of this redesigning is shown by means of PC experiments with different avenues regarding the ten biggest classes of the Reuters 21578.
Std16	In all tests, it was possible to download relevant snippets to enhance classification accuracy.
Std17	The results show an increase in the accuracy and detection rate and lower false caution rate.
Std18	The analysis shows the effectiveness of this technique in comparison with other state-of-the-art strategies.
Std19	The performance surpasses previous best results on benchmark datasets.
Std20	This hybrid approach accomplishes a 13% F-measure gain over previous rule-based approach and a 4% F-measure improvement vs. a manual classification process in a doctor's facility setting.

- **RQ7: Which journals have published these studies?**

Here, we see which journals publish ATC studies. Table 8. shows the distribution of articles according to science journals and other publications.

Table 8. The distribution of articles in data bases.

Database	Total result	Initial Selection	Final Selection
IEEE Xplore	**64**	**11**	**4**
Science Direct	**662**	**7**	**4**
Springer	**319**	**8**	**3**
ACM	**2547**	**4**	**3**
Scopus	**29**	**8**	**3**
Wiley	**246**	**4**	**3**
Total	**3867**	**42**	**20**

Figure 3. shows the distribution of articles according to science journals.

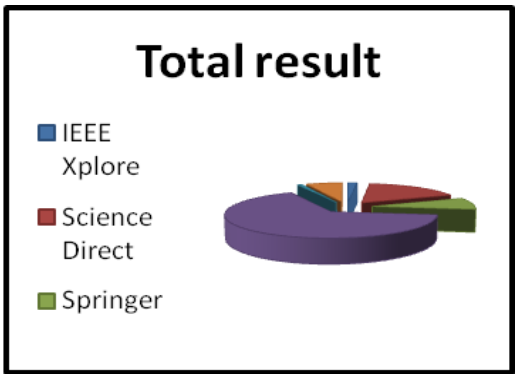


Figure 3. The distribution of articles.

- **RQ8: What are the acronyms used in the study manuscripts?**

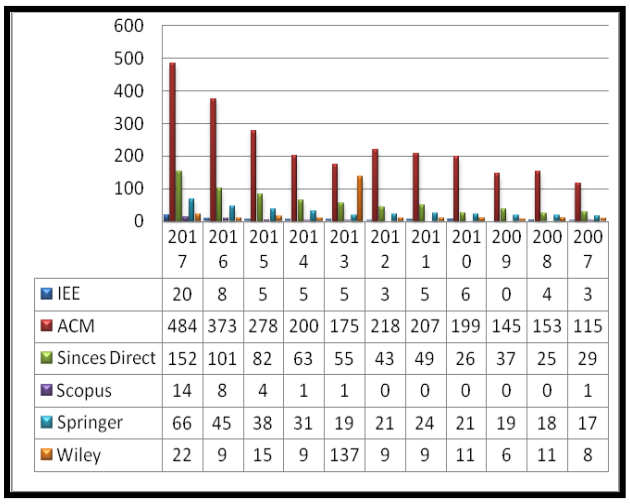
All contributions have many abbreviations, such as NLP (Natural Language Processing), ML (Machine Learning), and others. Table 9. shows the acronyms used in studies.

Table 9. The acronyms used in studies

Acronyms	Full Name	Articles
ELMR,AML,NB, KNN,	Evolutionary Learning of Meta-Rules, Automatic, Machine Learning, Multinomial Naive Bayes , K-Nearest	Std1
CNNs , LDA ,NLM ,MTI ,MESH.	Convolutional Neural Networks Latent Dirichlet Allocation, National Library of Medicine, Medical Text Indexer, Medical Subject Headings	Std2
IR,IE, C#.	Information Retrieval, information Extraction, , pronounced C Sharp.	Std3
SCA, *GT*,WWW.	Strong Covering Algorithms, *Good-Turing* , World Wide Web .	Std4
CBOW , WIPO, IPC , LSTM, NB, SVM, KNN.	Continuous Bag of Words, World Intellectual Patent Office, International Patent Classification, *Long Short Term Memory*, Naive Bayes Support Vector Machine, K-nearest	Std5
WSD, a-RS ,SSI.	Word Sense Disambiguation. adapted relation structure, Structural semantic interconnection,	Std6
VSM, kNN , SVM,TC , IR.	vector space model, K-Nearest Neighbors, Support Vector Machines, TEXT categorization, Information retrieval.	Std7
LSA GloVe, BOW.	Latent Semantic Analysis, Global Vectors, Continuous bag of words.	Std8
SVM, PWG.	Support Vector Machines, Anti-Phishing Work Group.	Std9
TC ,SVM	Text Categorization, Support Vector Machine.	Std10
IDF , VSM , IR ,TC ,IG.	Inverse document frequency, Vector space model, Information retrieval, Text categorization, Information gain.	Std11
NLP, CBOW, GloVe , MPQA	Natural language processing, Continuous Bag of Words, Global Vectors, multi-perspective	Std12
LDA ,IR ,BOW , PLSA. ,NLP	Latent Dirichlet Allocation , information retrieval, Bag-of-Words, Probabilistic Latent Semantic Analysis, Nature language processing	Std13
NLP ,HLBL	Nature language processing, Hierarchical log-bilinear	Std14
NB ,SVM ,LR ,C4.5	Naive Bayes classifier, support vector method, logistic regression ,Classification decision tree	Std15
NB ,SVM ,KNN	Naive Bayes Support Vector Machine, K-Nearest Neighbor.	Std16
IDS , C4.5 ,SVM.	Intrusion Detection Systems Decision tree , Support Vector Machine.	Std17
CNN ,BOW ,SVM	Convolutional neural network, Bag-of-Words, Support Vector Machine.	Std18
LSTM ,CNN	Long Short-Term Memory, Convolutional neural network	Std19
NLP/ML.	natural language processing plus Machine learning.	Std20

- **RQ9: Which years and journals have the highest number of articles related to the topic of our study?**

It appears in Figure 4. in 2017 and ACM journal have the highest numbers of publications, and this figure is explained the articles are published according to years and journals.



	2017	2016	2015	2014	2013	2012	2011	2010	2009	2008	2007
IEE	20	8	5	5	5	3	5	6	0	4	3
ACM	484	373	278	200	175	218	207	199	145	153	115
Sinces Direct	152	101	82	63	55	43	49	26	37	25	29
Scopus	14	8	4	1	1	0	0	0	0	0	1
Springer	66	45	38	31	19	21	24	21	19	18	17
Wiley	22	9	15	9	137	9	9	11	6	11	8

Figure 4. The schematic chart published papers

- **RQ10: Which models of text classification have been used in the studies?**

There are some models used in the studies subject to this paper which can be classified accordingly. Figure 5. explains the many models of text classification applied in this systematic mapping.

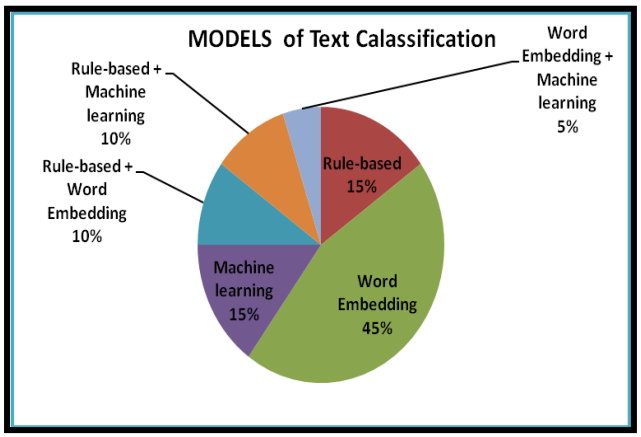


Figure 5. Models of text classification.

10. Addressing the Questions, Discussion, Conclusions, Future Work and Limitations of the Study

10.1. Addressing the Questions (ARQs)

This study contained ten research questions (RQ1 to RQ10), and in the following the main outcomes are listed.

❖ **RQ1: Types of studies**

Sometimes it is needed to implement simulations or experiments for results. Therefore, we have in this contribution two types of studies (empirical and experimental). A significant proportion (80%) of studies (16 out of 20) is contributed to experiments and only 20% is empirical.

RQ2: Domain areas of those papers

The studies are interested in many fields such as agencies, IT, electronics, and libraries. For this, there is a need to implement different programs related of classification techniques. Most contributions are in medical fields, patent, benchmark data collections, Internet banking, newspaper, Amazon and, finally, in the field of topic segmentation.

❖ **RQ3: Purpose of studies**

The aim of this question is to guide us to the focus of studies. For example, some studies discuss various methods for document representation, improving ML techniques, using a new rule-based method to detect phishing attacks on Internet banking websites, etc.

❖ **RQ4: The country of authors**

According to our study, with the articles having been chosen randomly and considering word embedding, ML and Rule-based as part of ATC techniques, we can see USA, China, and India in the first level, Singapore and Mexico in second, and others such as Russia, Belgium, Italy, Tunisia, Canada, and the rest in subsequent positions.

❖ **RQ5: The types of models**

In classifications, different methods are used like Word2Vec, GloVe, RIPPER, C4.5, Naïve Bayes, and SVM. It is clear that SVM has 40% of all models in 20 contributions, TF-IDF 10% and C4.5 10%.

❖ **RQ6: The validity criteria**

The validity of studies can be seen by depending on the performance of the classification of systems and also on different metrics such as, accuracy, precision and recall. For this reason, the best study will be regarded as having the highest validity. For example, the analyses in Std.9 show that the introduced display by utilizing the proposed feature sets along with some relevant features can detect phishing pages in Internet banking with the accuracy of 99.14% true positive and only 0.86% false negative alarm.

❖ **RQ7: Main journals**

By scanning the science citation index, it is shown that ACM has the largest number of articles equal to 2547, and IEEE has less number of articles equal to 64 from the total number (3867).

❖ **RQ8: The Acronyms**

All studies have many abbreviations, such as, IDF, VSM, IR, TC, IG....etc. So, NLP (Natural Language Processing), ML (Machine Learning) and ATC (Automatic Text Classification) are the most common ones.

❖ **RQ9: Journals and years of publication**

The year 2017 and the ACM journal have the highest number of publications.

❖ **RQ10: Models of Text Classification**

Word-embedding has 45%, ML 15%, Rule-based 15%, Rule-based and word-embedding together 10%, Rule-based and ML together 10% and word-embedding and ML together have 5% of all the models used for text classification.

10.2. Discussion

This section presents an interpretation of the results of this systematic mapping study and the implications of the results for researchers, and we will explain in next paragraphs the results of this study.

- This study is an attempt to determine the areas for Word embedding, ML and ATC as in different departments such as computer science, and IT.
- This work contributes to many global and scientific fields for using Word embedding, ML and ATC such as:
  - ✓ Medical field
  - ✓ Classification of text documents in news agencies, such as Reuters's Dataset, The Anatolia news agency, etc.
  - ✓ Electronic Libraries, where it becomes even more important since millions of electronic books exist in many Internet sites.
- It can be stated that Word embedding, ML, and ATC, which were covered in our research community as both empirical and experimental, and those experimental studies have an area greater than empirical ones.
- Some studies employ one, two more methods of machine learning algorithms, such as Naïve Bayes, SVM.....etc., for comparison to choose the best alternative.
- It is a necessity to make an assessment of each study to determine their quality, and validity is one of such assessment and measurement metrics.
- It is more important to follow systematic ways to analyze documents and to make their classification easier by applying specific methodologies.
- Finally, it is shown that systematic mapping study is the most beneficial when attempting to obtain summaries of studies, especially in ATC, and that it is better than Systematic Literature Review (SRL).

10.3. Conclusion

The main motivation for this work is to investigate the state of Word embedding, ML and ATC by systematic mapping in order to determine what issues have been studied, as well as by what means, and to provide a guide to aid researchers in planning future research. The mapping Study used here is beneficial in identifying the files, where Word embedding, ML, and ATC are the most effective and prominent as well as those areas where more research is needed. Moreover, upon research through the literature, some important aspects are found to have not been reported, and in other cases, only a brief overview is given. In addition, regarding industrial experiences, the authors notice that they are rare in the literature. The present systematic mapping study provides a structure of the type of research reports

and views that have been made so far in a more comprehensive way. This project included 20 articles in many fields, where different types of algorithms have been used with satisfactory results and attempted to enhance the automatic text classification approach into wider and more global fields.

10.4. Future work

We can adopt a new approach by inserting additional questions to get a comprehensive view of other researches.
Real-time can be applied in the systematic mapping study approach.
Other approaches such as the mixing images and text can be used in future to classify documents with pictures and text.

10.5. Limitations of study

The validity of a mapping study depends on similarities in primary studies [22, 23]. In any case, the list of studies might not have been entirely complete and may have had their own restrictions. As a result, additional or arbitrary terms, for example, 'system quality', may have changed the final list of the papers examined [24]. Additionally, the references in the selected studies were not checked to identify other related works, so it is represented the threats to validity, this validity is important to judge the strengths and limitations of our systematic mapping study. Finally, six databases were used in our systematic mapping study (Science Direct, ACM, IEEExplorer, Springer Link, Scopus and Wiley) because they represented the most important and relevant databases for the aim of this study. In future works, other databases may also be examined.
For our study, the following issues may be represented as threats to validity:

- Researcher bias with regards to exclusion/inclusion.
- Selection of search databases.
- Search terms and time frame.
- Data extraction (classification).

The standard classification scheme of validity threats was suggested by Wohlin et al. (2000), it was used in this study, and it can discuss the issues above in relation to four types of threats to validity:

1. Conclusion validity.
2. Construct validity.
3. Internal validity.
4. External validity.

As a result, the final decision to select a study depended on some authors who conducted the search process.
In this respect, the main limitations affecting our study are:

10.5.1. Conclusion validity:

It refers to the degree to which conclusions we reach about the relationships are reasonable. However, the conclusion drawn about the research view in this field includes academic trends and the subjects that have been studied or discussed. These conclusions depend on statistical data from paper datasets and focused on a specific issue such as article facet, research facet, etc. The conclusion validity issue lies in whether there is a relationship between the actual academic focus and the number of articles. When that relationship does not exist, validity is compromised.

10.5.2. Internal validity:

It deals with extraction and analysis of data [23]. Yet, classification of the primary studies is implemented by some authors, while the review of the final results is done by others.

10.5.3. Construct validity:

It is concentrated on the relationship between the theory that experiments depend on and the observations related to them. However, construct validity is for selecting the right variables to compute the phenomenon of interest. Construct validity in our contribution lies in the comprehensiveness of the classification scheme used for the data extraction.

10.5.4. External validity:

It is about the generalization of our study [25]. In the selected study, the papers were chosen as written in the English language, and articles written in other languages were excluded, such as (Arabic, Spanish, Chinese, etc.). However, one matter lies in that whether the articles included in our database represent all the relevant works in the field of Automatic Text Categorization, Rule-based and ML.

References

[1]. McConnell, S. (2008). Managing technical debt. *Construx Software Builders, Inc*, 1-14.

[2]. Dybå, T., Kampenes, V. B., & Sjøberg, D. I. (2006). A systematic review of statistical power in software engineering experiments. *Information and Software Technology*, *48*(8), 745-755.

[3]. Hannay, J. E., Sjoberg, D. I., & Dyba, T. (2007). A systematic review of theory use in software engineering experiments. *IEEE transactions on Software Engineering*, *33*(2), 87-107.

[4]. Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, *10*(1), 1-309.

[5]. Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, *3*(Feb), 1137-1155.

[6]. Ligeza, A. (2006). *Logical foundations for rule-based systems*(Vol. 11). Heidelberg: Springer.

[7]. Durkin, J., & Durkin, J. (1994). *Expert systems: design and development* (pp. 1-800). New York: Macmillan.

[8]. Durkin, J.(1996) " Expert Systems: Catalog of Applications". Intelligent Computer Systems, Inc, Journal IEEE Expert: Intelligent Systems and Their Applications , Vol 11 Issue 2, Pages 56-63,April .

[9]. Nikolopoulos, C. (1997). *Expert systems: introduction to first and second generation and hybrid knowledge based systems*. Marcel Dekker, Inc..

[10]. Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

[11]. Antonie, M. L., & Zaiane, O. R. (2002). Text document categorization by term association. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on* (pp. 19-26). IEEE.

[12]. Buddeewong, S., & Kreesuradej, W. (2005, November). A new association rule-based text classifier algorithm. In *Tools with Artificial Intelligence, 2005. ICTAI 05. 17th IEEE International Conference on* (pp. 2-pp). IEEE.

[13]. D Maghesh Kumar (2010),"Automatic Induction of Rule Based Text Classification", International Journal of Computer Science and Information Technology (IJCSIT), Vol.2, No.6.

[14]. Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

[15]. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929-1958.

[16]. Zeiler, M. D. (2012). ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*..

[17]. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

[18]. Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, *12*(Jul), 2121-2159.

[19]. Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering–a systematic literature review. *Information and software technology*, *51*(1), 7-15.

[20]. B. Kitchenham, S. Charters .(2007). Guidelines for performing Systematic Literature Reviews in Software Engineering, Technical Report EBSE 2007-001, Keele University and Durham University Joint Report.

[21]. Petersen, K., Feldt, R., Mujtaba, S., & Mattsson, M. (2008, June). Systematic Mapping Studies in Software Engineering. In *EASE* (Vol. 8, pp. 68-77).

[22]. Ampatzoglou, A., Charalampidou, S., & Stamelos, I. (2013). Research state of the art on GoF design patterns: A mapping study. *Journal of Systems and Software*, *86*(7), 1945-1964.

[23]. Elberzhager, F., Münch, J., & Nha, V. T. N. (2012). A systematic mapping study on the combination of static and dynamic quality assurance techniques. *Information and software technology*, *54*(1), 1-15.

[24]. Garousi, V., Mesbah, A., Betin-Can, A., & Mirshokraie, S. (2013). A systematic mapping study of web application testing. *Information and Software Technology*, *55*(8), 1374-1396.

[25]. Easterbrook, S., Singer, J., Storey, M. A., & Damian, D. (2008). Selecting empirical methods for software engineering research. In *Guide to advanced empirical software engineering* (pp. 285-311). Springer, London.

The Studies

Sd1. Gomez, J. C., Hoskens, S., & Moens, M. F. (2017, July). Evolutionary learning of meta-rules for text classification. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion* (pp. 131-132). ACM.

Sd2. Xu, H., Dong, M., Zhu, D., Kotov, A., Carcone, A. I., & Naar-King, S. (2016, October). Text classification with topic-based word embedding and convolutional neural networks. In *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*(pp. 88-97). ACM.

Sd3. Gowri, S. C., & Sundaram, D. K. M. (2015). A Study on Information Retrieval and Extraction for Text Data Words using Data Mining Classifier. *International Journal of Computer Science and Mobile Computing*, *4*(10), 121–126.

Sd4. Liu, J., & Lu, Y. (2007, August). An Ensemble Text Classification Model Combining Strong Rules and N-Gram. In *Natural Computation, 2007. ICNC 2007. Third International Conference on* (Vol. 3, pp. 535-539). IEEE.

Sd5. Grawe, M. F., Martins, C. A., & Bonfante, A. G. (2017, December). Automated Patent Classification Using Word Embedding. In *Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on* (pp. 408-411). IEEE.

Sd6. Dhopavkar, G., Kshirsagar, M., & Malik, L. (2015, March). Application of Rule Based approach to Word Sense Disambiguation of Marathi Language text. In *Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015 International Conference on* (pp. 1-5). IEEE.

Sd7. Lan, M., Tan, C. L., Su, J., & Lu, Y. (2009). Supervised and traditional term weighting methods for automatic text categorization. *IEEE transactions on pattern analysis and machine intelligence*, *31*(4), 721-735.

Sd8. Naili, M., Chaibi, A. H., & Ghezala, H. H. B. (2017). Comparative study of word embedding methods in topic segmentation. *Procedia Computer Science*, *112*, 340-349.

Sd9. Moghimi, M., & Varjani, A. Y. (2016). New rule-based phishing detection method. *Expert systems with applications*, *53*, 231-242.

Sd10. Marseguerra, M. (2014). Early detection of gradual concept drifts by text categorization and Support Vector Machine techniques: The TRIO algorithm. Reliability Engineering & System Safety, 129, 1-9..

Sd11. Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF* IDF, LSI and multi-words for text classification. Expert Systems with Applications, 38(3), 2758-2765.

Sd12. Li, Y., Pan, Q., Yang, T., Wang, S., Tang, J., & Cambria, E. (2017). Learning word representations for sentiment analysis. *Cognitive Computation*, *9*(6), 843-851.

Sd13. Niu, L. Q., & Dai, X. Y. (2015). Topic2Vec: learning distributed representations of topics. *arXiv preprint arXiv:1506.08422*.

Sd14. Turian, J., Ratinov, L., & Bengio, Y. (2010, July). Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 384-394). Association for Computational Linguistics.

Sd15. Gulin, V. V., & Frolov, A. B. (2016). Categorization of text documents taking into account some structural features. *Journal of Computer and Systems Sciences International*, *55*(1), 96-105.

Sd16. Guzmán-Cabrera, R., Montes-y-Gómez, M., Rosso, P., & Villasenor-Pineda, L. (2009). Using the Web as corpus for self-training text categorization. *Information Retrieval*, *12*(3), 400-415.

Sd17. Koshal, J., & Bag, M. (2012). Cascading of C4. 5 decision tree and support vector machine for rule based intrusion detection system. *International Journal of Computer Network and Information Security*, *4*(8), 8.

Sd18. Johnson, R., & Zhang, T. (2014). Effective use of word order for text categorization with convolutional neural networks. *arXiv preprint arXiv:1412.1058*.