Journal of Theoretical and Applied Information Technology
15th January 2022. Vol.100. No 1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DOCUMENT CLASSIFICATION SYSTEM FOR THE SPANISH LANGUAGE

LUIS GABRIEL MORENO SANDOVAL[1], LILIANA MARIA PANTOJA ROJAS [1 2], NELSON GIOVANNI AGUDELO CRISTANCHO [3], CRISTINA RAMÍREZ MENESES [3]

[1] GICOGE Research Group, Universidad Distrital Francisco José de Caldas, Colombia

[2] LUMON LV TECH Research group, LUMON SAS, Colombia

[3] SUOMAYA CGMLTI SENA Research Group, SENA, Colombia

E-mail: gabriel.moreno@lumon.com.co

ABSTRACT

The classification of documents is a relevant task in companies to save time in managing information present in specific documents; therefore, the health sector seeks to prioritize documents performing the traceability of any process within its network. This article presents a document classification system to provide a tool divided in software components that faces the challenges of binding to the Spanish language using public sources such as Google and Wikipedia applying long documents related to the health sector in Colombia. For this purpose, a set of Machine Learning classifiers is performed to compare F1-score, Precision, and Recall metrics obtaining the best performance in the Logistic Regression classifier. In addition, the article makes a theoretical survey on the relationships that text mining, Information Retrieval, and Text Summarization have with document classification.

Keywords: *Text Mining, Information Retrieval, Text Summarization, Document Classification, Spanish Language.*

1. INTRODUCTION

The classification of documents is one of the main tasks that has been carried out as a result of the development and study of Natural Language Processing and Computational Linguistics. This article's main objective is to propose an effective classification system to face the challenges and complexities of the Spanish language by collecting documents from public sources. In this way, the classifier becomes a tool to establish valuable information bases that allow informed decision-making by people in management positions in both public and private institutions to take advantage of the content of thousands and thousands of archived or digitized physical documents.

Therefore, to perform the correct development of this main task, the article collects some geolects through the particular documents that are managed in Colombia, creating a dataset divided into training and testing through public sources such as social networks (Google and Wikipedia) to obtain the best model among a set of five text classifiers belonging to the field of Machine Learning. Thus, the documents must be converted into an input for an unstructured data analytics tool that can take advantage of the transversality of the methods of related tasks such as Text Mining (TM), Information Retrieval (IR), and Text Summarization (TS), leaving the contributions in this transversality for future research. Additionally, it should notice that a big part of the algorithms' development for this task and those related are configured in languages like English, leaving a considerable gap when the classification analysis is in Spanish.

The paper presented is structured as follows: the first section presents the literature review of the fields related to document classification along with some studies demonstrating the relationship between these and the task of interest; the second section shows the methodology describing the binding processes, hyperparameters used in the classifier and the selected dataset. The next section presents the results to demonstrate the classification performance

executed with the data set described above, adding a final discussion framed in the critical evaluation in the light of the works cited and explanation of conflicts in the literature. Finally, the conclusions analyze the document classification by mentioning the current knowledge's limitations, describing the importance of the work in the field, and presenting future work.

2. LITERATURE REVIEWS

Text extraction is one of the challenges of text mining to provide valuable insights into available information to generate added value. Han et al. [1] referred to text mining as an interdisciplinary field based on Information Retrieval (IR), Data Mining (DM), Machine Learning (ML), statistics, and Computational Linguistics (LC); consequently, advanced text mining techniques have been implemented to applications in new fields to solve challenges in the face of large unstructured data processing capacity.

The following sections present the theoretical background of the challenges related to the main task of this article: document classification, where each developed section highlights the contributions in the field of Natural Language Processing (NLP) and the relation of their contributions to the main task.

2.1 Text mining

Text mining is based on several advanced techniques from statistics, ML, and linguistics. Its goal is to be able to "*process large textual data to extract high quality information, which will be useful to provide insights into the specific scenario to which it is applied*" [2]. Therefore, its main task is to look for patterns and trends in unstructured data, which has allowed it to have many uses, including text clustering, concept extraction, sentiment analysis (SA), and TS.

Its first applications arose from the need to organize documents [3], so that a new possibility of creating summaries with descriptive statistics was born, making it possible to take advantage of the large volumes of documents deposited in filing cabinets. With time and new scientific fields, text mining techniques for analyzing text developed as an automatic, dynamic, and inferential process were sought. Hearst [4] recognized such developments: "*a mixture of computer-driven and user-driven analysis*" (p. 8), which is the core of today's supervised models used to perform predictive analysis, a superior process of analysis for harnessing information from their digitized counterparts.

Perhaps the most famous text mining applications for AI generation are IBM's Watson, Amazon's Alexa, Microsoft's Cortana, Apple's Siri, and Google Assistant. For example, Kotu and Deshpande [5] mentioned Watson's outstanding performance "*the spectacular performance it had, when it competed against humans on the late-night game show Jeopardy*" (p. 282). These authors and others allude to Watson's excellent performance because it can access large amounts of structured and unstructured data from comprehensive databases, such as Wikipedia.

As we find ourselves in a world with multiple languages, other cases of analysis are involved, such as multilingual data mining, multidimensional text analysis, contextual text mining, and the evolution in text data with applications to particular fields of work: text mining in security, biomedical literature analysis, online media analysis, and analytical customer relationship management. Han, et al. [1] named some of the types of software, analysis, and text mining tools in academic institutions, open-source forums, and industry dedicated to the implementation of multiple languages as are: WordNet, Semantic Web, Wikipedia, and other information sources that enable improved understanding and extraction of text data.

2.1.1 Text mining and document classification

The significant growth of information and electronic documents that we can find in the World Wide Web requires a correct classification of all this information so that organizations can perform the correct interpretation. However, performing this manual task requires a lot of time and concentration, thus turning document classification into an area of interest for data mining generating different techniques such as Linear Discriminant Analysis (LDA), Naïve-Bayes, K-Nearest Neighbors, Artificial Neural Networks (ANN), among others [6].

Classification of documents is a drawback for text mining, characterized by the words generating a dataset in the document, cluttering the document, and thus the context of the document [7]. Therefore, some research has focused on the sentiment in documents to classify them and

demonstrate this bridging relationship between text mining and document classification. For example, Tang et al. [8] incorporated user-product level information about comments from the IMDB and Yelp dataset into a neural network model under a supervised learning approach producing better classifications when making text representations to capture sentiment better. However, these models present a challenge when documents are large due to the insufficient memory unit; Xu et al. [9] proposed a cache mechanism that divides memory within groups with different forgetting rates of information to keep sentiment captured from a recurring unit assimilating human brain behaviors.

Similarly, Wen et al. [10] used the same datasets proposing a speculative sentiment classification model, which added to Tang's approach two components. The first is document coding using word embedding resulting from a learning process about products. The other is based on the hypothesis that users with the same classification behavior are likely to write a similar document about a product, thus generating the speculative similar document component. On the other hand, Yang et al. [11] approach documents' hierarchical structure under an attention mechanism of document words or phrases. They produce visualization layers by aggregating the words with the highest levels of importance to build relevant phrase vectors and thereby demonstrate qualitatively informative phrases that improve document classification models.

2.2 Information Retrieval (IR)

It is a field of scientific findings and practices paired with DM, whose goal is to collect, manage, process, analyze, and visualize a vast amount of structured and unstructured [12] information presented in the text. Nowadays, IR systems work with search engines and are standard such as Google, PubMed, Apple's Spotlight. For example, in biomedical sciences, PubMed queries a database of 21 million scientific publications from the National Library of Medicine [13].

Consequently, the usual models in this field are designed to discover text patterns using external sources such as thesauri and ontologies that limit text extraction to patterns on a specific domain of the topics contained in the texts. However, Atkinson et al. [14] described the beginning of the inclusion of Genetic Algorithms (GAs), allowing to take advantage of their ability to: develop a global search, the exploration of solutions in parallel, the robustness to solve problems of missing or biased information and the ability to evaluate the solution likelihood.

Also, the authors Pérez and Codina [15] examined the usual architecture within the information retrieval systems, which has two components: the indexing system and the query system, where the first analyzes the data that is downloaded from the web to subsequently establish indexes that will help to perform searches, on the other hand, the second is where it interacts with users, better known as the search engine interface.

Tenopir et al. [16] showed that the study of mental models and learning styles in college students using the web for information retrieval systems had a relationship. It was found that individuals' mental models affect their search results and their feelings towards searching, thus concluding that the construction of individuals' mental models on the web is linked to their experiences, personal observations on the searched topic, and classroom instruction.

It is observed how users' mental models of databases are based on their perceptions of information systems and are modified with ongoing experience [17]. Zhang [17] studied the users' interaction with Medline Plus, suggesting that individuals' development of a mental model of a system evolved from their experiences with another system. According to the author, this focused on "assimilating new elements into mental models and modifying existing elements" (p. 168).

2.2.1 Information retrieval and document classification.

Organizations need to have correct document management; hence with the increase of information in digital form, traditional methods are not relevant to meet the needs of the same organizations. Automated technology allows classifying and retrieving documents; its technology is used to compare all documents by performing queries using keywords. When incorporating classification with document retrieval, it becomes a valuable tool because having established the critical characteristics of each document, only the documents that are highly related to the established characteristics will be retrieved, thus allowing a high retrieval speed and greater accuracy. Additionally, for any company, the classification of documents facilitates the

processes and activities to be performed, also retrieves similar documents found in the database [18].

A clear example is the R+D+I industry, where usually the documents are digital, and large quantities of documents are stored, making necessary keywords that serve as criteria allowing a better location of the documents. Therefore, it is necessary to generate a prototype that allows automatic characterization because being manually is influenced by human factors. For this reason, the author Lin [19] generated a prototype that allows the classification and retrieval of documents, being this hybrid model and based on SVM, to facilitate these activities in the R+D+i industry. This prototype was tested in the semiconductor industry in Taiwan, which at the time of being installed and performing pilot tests gave promising results such as efficiency in document management, having a document retrieval time of minutes to seconds; can be concluded that the phases of this process were: expansion and adjustments of the knowledge dimension categories, automatic classification technology, and connections.[19]

The authors Khalifi et al. [20] proposed a formal model and a search algorithm with which they expected to find features of the information items and subsequently structure the search results. In such a way, they used the following stages: NLP, statistical representation of queries and documents, and an autonomous learning model in order to determine the most relevant results; all this was performed using a Yahoo database, where the system reflected significantly satisfactory results.

2.3 Text Summarization

Text summarization is an essential task of NLP, which follows in several applications. The two broad categories of text summarization approaches are extraction and abstraction. *"Extractive methods select a subset of existing words, phrases, or sentences from the original text to form a summary"* [21] Although this would be meaningless if the context is not applied to generate value; thus, author Song et al. [22] posited that "*Text summarization involves the production of text summaries that incorporate the essential information of the text(s), taking into account the readers' interests*." (p. 125).

Currently, search engines have established a text summary when displaying search results to give a brief description of such results. In the case of Google, this presents a short text summary of an essential item placing it at the top of the list of results. The details of how this process is performed are not yet in the public domain, but Zhuge [23] stated that the text summary is likely to be generated using keywords from the phrase entered on the web page, identifying the search purpose and thus organizing the information for the user.

Another classic example of text summarization is the conversion of text represented in an image. Zhuge [23] indicated that one solution is to transform this issue into information retrieval and text summarization problem: Select the representative tags of these images, search for relevant texts according to these tags (or select texts that contain or link to these images), and then summarize these texts. Another way is to establish essential semantic links between the images, find the text that best fits, and make the necessary text summary. The establishment of semantic links is based on the relationships between their tags determined by existing texts, semantic links, and image categories.

For example, one of the essential fields where these techniques are applied is research in the field of biology, which has suffered from inconsistent descriptions of gene products and ambiguous definitions of terms from disparate biology databases, which also hinders computational semantic processing of biomedical literature such as text summarization or document clustering. To this end, ontology would be a promising solution to such conflicts of inconsistent descriptions. Currently, no ontology captures the full range of concepts in the biomedical domain; however, Karp [24] mentions the existence of several well-designed biomedical ontologies, such as the UMLS (Unified Medical Language System), the Gene Ontology (GO), and the EcoCyc Ontology.

The evolution of methodologies has taken advantage of using these ontologies mixing several applications such as text mining, spam filtering, text summarization, and bioinformatics, where it is vital to quantify the "similarity" of two strings. However, Theodoridis [25] saw a problem in the common use of kernels for this problem; by their definition, they are similarity measures; they are constructed to express inner products in high-dimensional feature space. An inner product is a similarity measure, and hence two vectors are more similar if they point in the same direction. Thus, there has been a great deal of activity in defining kernels that measure similarity between strings from this observation.

2.3.1 Text summarization and document classification

Text summarization is seen as a task that deals with the semantics of the analyzed text, providing a different field of analysis to classification problems. For example, Mamakis et al. [26] overcome some supervised and unsupervised approaches with a methodology that assists text summarization results obtained from a classification model, demonstrating the feasibility between these tasks. Furthermore, they recommend using spatial features to improve summarization performance on a dataset such as newspaper articles. Similarly, Kumar et al. [27] present a hybrid model showing its advantages after this approach using the Document Understanding Conference corpus to perform a text summarization model that filters high-quality sentences from a classification model. The study produced better classification performance and alleviated some limitations of classical text summarization models.

Among the advances for establishing the relationship of these tasks, an increased performance from performing hybrid models has been demonstrated. Thanh et al. [28] perform a text summarization combining the searching maximal frequent word sets and clustering algorithms with which they will subsequently create the feature vector to minimize the dimension of the vector over the documents that will be the subject of the classification models; thus outperforming approaches with supervised learning model. With the same approach, Dewi and Sagala [29] perform the text summaries on 100 thesis documents to classify the resulting documents. However, they showed that some techniques such as TF-IDF could not be used when performing these types of hybrid models.

2.4 Document Classification

Document classification is an autonomous learning algorithm where the words that appear in the document are characterized; that is, this task performs according to the presence or absence of each word. Automatic document classification is a topic that has been gaining momentum worldwide, making it a significant area of research. In this way, large volumes of documents that support processes within companies or research records are converted into bytes of information processed by computers with significant memories.

The traditional classification processes for document management are carried out according to the textual analysis of the content of the documents and consist of two characterizations. Witten and Frank [30] mention such characterizations: the first one consists of defining a Boolean attribute; on the other hand, the second characterization defines a Bag of Words. As the importance of the words is granted according to the number of times they are repeated supported with the characterizations, we move on to solve a computational time optimization problem under the classification methodology employed to perform this task.

The process of document classification can be performed under two methodologies: supervised learning, which has the categories known in advance for each document; on the other hand, there is unsupervised learning where the categories or names of the classes are not defined, but it searches for groups of related documents allowing in this way to generate links with other documents that are considered relevant for a query [30].

According to the analysis made by Khan et al. [31] determined that chi-square statistics were a good method for feature selection and classification used; but it is more appropriate to use hybrid methods, among these, we find the Super Vector Machine (SVM), Naive Bayes (NB) and k-Nearest Neighbours algorithms. Although they are more appropriate, these authors also mention that such methods in the text mining aspect are insufficient because the semantics of the documents require a deeper investigation. Consequently, ontology and semantics are born as additional resources that improve the accuracy and qualification process.

There are different techniques such as LDA, Naïve-Bayes, K-nearest neighbors, ANN, among others. One of the frequently used techniques for document classification is the Naïve-Bayes algorithm because this technique is fast, accurate, and simple, achieving impressive results and often outperforms more sophisticated classifiers [30]. This method generates a set containing all the unique words in the text by using probability (independent of the context and the word's position in the document). Subsequently, derivations to this method were developed as the assumptions were relaxed; Multinomial Naive Bayes is an example that considers the number of occurrences of each of the words.

Additionally, a study conducted in the United Kingdom on students at the General Certificate of Secondary Education (GCSE) focused on automatically classifying documents with short answer questions. Initially, they applied standard

data mining techniques to the model corpus in order to measure the similarity between the student's answers and the model corpus, based on the number of common words and the use of k-means clustering. The study concluded a considerable gap between students' responses and the model corpus, which was reflected in the fact that the number of words used correctly influences the scores more than the semantics or word order set in the model corpus [32].

On the other hand, it is vital to compare semantic text classification and traditional text classification based on the five approaches of Altınel and Ganiz [33]: domain knowledge, corpus, deep learning, word or character sequence, and enriched linguistics. They concluded that it is difficult and costly to process a large external knowledge database for both classifications when it is a deficient lexical database because these are in a limited number of languages. Otherwise, in some classification algorithms, it is possible to understand and explain the model and the decision analysis obtained by the model when the researchers applied different data mining techniques on knowledge bases and corpus-based systems.

It should also be noted that some document classification methods have progressed with the use of keywords to deal with the presence of ambiguity. However, because a word has different meanings and multiple words can have the same meaning, the authors Liu et al. [34] proposed to develop an automatic text classification method based on disambiguation utilizing WordNet with the hierarchy of similar concepts through Synset and applying the Brown Corpus. Furthermore, they compare the text classification by manual disambiguation offered by Princeton University, with which it is intended to achieve contributions in the management systems of web pages, digital books, digital libraries, among others.

Due to the boom presented by document classification and advances in natural language processing (NLP) and text mining, many researchers have focused on developing four phases: feature extraction, dimension reductions, classification algorithms, and evaluation. Therefore, authors Kowsari et al. [35] considered that understanding text classification methods and their correct evaluation is relevant for decision making. Therefore, they established steps to refine the mentioned phases; the first phase determines that text and document cleaning is essential to improve accuracy and robustness; the second phase allows to optimize the time and memory complexity of existing algorithms in document classification. The third phase should use the algorithm appropriate to the expected results, and the fourth phase indicates the metrics such as the coefficient Matthews correlation coefficient (MCC), Receiver Operating Characteristic (ROC), and Area Under the Curve (AUC) that allow to evaluation the document classification algorithm.

Finally, this field has expanded so that methods are implemented to perform document classification when the content of the documents are images. This shift in focus is on identifying and finding a researcher-determined figure within an image [36]. Like its text analysis counterpart, the medical field has also used this approach by training thousands of images to support disease diagnosis so that detection is automatic

3. METHODOLOGY

To achieve the article's objective by developing a document classifier in the Spanish language, Figure 1 represents the three components of the methodology.

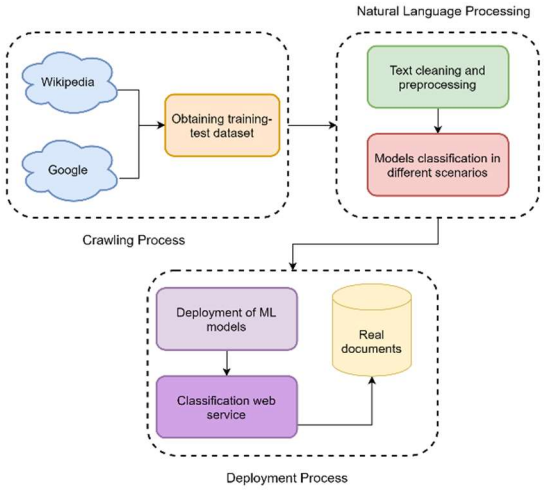

Figure 1 Pipeline Classification Process

The first component consists of performing a crawling process in the digital social networks of Wikipedia and Google in order to take advantage of public sources. This process performs searches for the definitions of the classes in Spanish to automatically obtain those that will be the object of the classification models; in this case, the crawling process was configured to obtain a set of 100 documents per class. Subsequently, the obtained documents are organized in such a way as to obtain

a dataset that will divide into subsets of training and test data.

The second component describes the PLN process on the text of each document obtained; consequently, the first step is to perform text cleaning and preprocessing on all documents. In each document, URL and special punctuation characters are eliminated, w where the characters of the text are converted into lowercase. Then, the paragraphs are tokenized to eliminate the Spanish language stopwords and those stored in a Bag of Words that do not contribute to some encoding of the text that does not make the text understandable. Finally, stemming and lemmatization are performed to find these base words making a heuristic process that eliminates the derivational affixes of these and a morphological analysis that eliminates the inflectional endings of the words. In addition, the classification models are trained by configuring their hyperparameters that will guide the model to perform the classification correctly.

The third component puts the classification models into production by validating the training performances by deploying the models which use ML. Then, these models are connected to the classification web service to produce performance statistics by placing the actual documents that refer to a set base external to the training.

3.1 Dataset

The dataset contains the documents obtained in Spanish from the digital social networks Google and Wikipedia to define the words that each type of document should have, thus indicating the classes of the classifiers. Table 1 describes the number of documents obtained from both social networks for the five classes: Bill, Tutela action, Medicines, Health system in Colombia, and Mandatory Health Plan.

Table 1: Dataset

Class	Document type	# documents
0	Bill	98
1	Tutela action	98
2	Medicines	100
3	Health system in Colombia	88
4	Mandatory health plan	99
Total		**483**

The following concepts are the class definitions of classification models to understand the content of the documents.

- Bill: according to the Colombian commerce code, the bill is a:

  "Value title that the seller or service provider may issue and deliver or remit to the buyer or beneficiary of the service."

- Tutela action: according to the Colombian ombudsman's office, these documents are understood as:

  "It is a constitutional mechanism for protecting human rights that empowers any person to go before a judge at any time or place to seek a pronouncement that protects a fundamental constitutional right. A person claims when the right is violated or threatened by the action or omission of public authorities or individuals in the cases determined by law, provided that there is no other suitable judicial defense mechanism unless it is to avoid irretrievable damage, event in which it proceeds as a transitory mechanism.

- Medicines: The Colombian Medicine Pricing System defines medicine as:

  An obtained pharmaceutical preparation from active principles, with or without auxiliary substances, is presented under the pharmaceutical form to prevent, relieve, diagnose, treat, cure, or rehabilitate the disease. Containers, labels, tags, and packaging are an integral part of the medicine since they guarantee quality, stability, and adequate use.

- Health system in Colombia: Guerrero et al. [37] describes the Colombian health system as follows:

  The Colombian health system is composed of a broad social security sector financed with public resources and a decreasing private sector. Its central axis is the General Social Security Health System (GSSHS). Affiliation to the system is mandatory and is done through public or private Health Promoting Entities (HPE), which receive the contributions and, through the Service Providing Institutions (SPI), offer the Obligatory Health Plan (OHP) or the OHP-S for those affiliated to the Subsidized Regime (SR). (p.3)

- Mandatory Health Plan: The Ministry of Health

and Social Protection defines the mandatory plan in the following words:

> *The backbone of the GSSHS in Colombia is the OHP, which corresponds to the package of essential services in the areas of health recovery, disease prevention, and income coverage on a transitory basis - economic benefits - when there is an inability to work due to illness, accidents or maternity.*

The models were divided into training and test subsets; the first subset for training has 60% of the samples, and another one has 40% of the samples. Table 2 shows how the documents obtained in the previous phase were divided, fitting each class into a partition close to 60/40. The results were developed with a test subset along with all the classification models.

Table 2: Training and test dataset

Class	Training	Test
0	56	42
1	59	39
2	61	39
3	52	36
4	61	38
Total	**289**	**194**

4. RESULTS AND DISCUSSION

Different classification models were programmed for the texts with the scikit-learn library, such as Logistic Regression (LR), Naive Bayes Multinomial (NBM), Complementary Naive Bayes (CNB), Random Forest (RF), and Super Vector Machine (SVM). Table 3 describes the configured parameters of the classification models; therefore, the other parameters of these models were set by default. Thus, all parameters for the NBM and CBN models are set to default.

Table 3: Classification model parameters

LR		RF		SVM	
Parameter	Value	Parameter	Value	Parameter	Value
Random number generator used (random_state)	0	Number of trees in the forest (n_estimators)	200	Kernel type (kernel)	linear
Algorithm in the optimization problem (solver)	lbfgs	The maximum depth of the tree (max_depth)	6	Regularization parameter (c)	0.5
Approach for handling multiple classes (multi_class)	multinomial	Randomness of the bootstrapping of the samples (random state)	0		

Next, Tables 4 - 8 demonstrate the classifier's performance under the three metrics of Accuracy, Recall and F1-score for each class included. Table 4 shows the performance for the LR classifier, which shows an overall level above 79% of correctly classified documents highlighting the medicine class with this precision metric. Likewise, Recall shows a good level of correctly identified documents with a minimum rate of 74%; however, the class with the best performance under this metric is the tutela action.

Table 4: LR classifier performance

Class	Precision	Recall	F1-score
0	0.79	0.74	0.77
1	0.91	1.00	0.95
2	1.00	0.97	0.99
3	0.72	0.78	0.75
4	0.80	0.74	0.77
Accuracy			**0.85**
Macro average	**0.84**	**0.85**	**0.84**
Weighted average	**0.85**	**0.85**	**0.84**

Table 5 shows the performance for the NBM classifier, demonstrating a wide range depending on the class to be analyzed in the correctly classified documents. However, mandatory health plan documents are the second-best identified class. Finally, medicine is the best performing class in both metrics.

Table 5: NBM classifier performance

Class	Precision	Recall	F1-score
0	0.84	0.64	0.73
1	0.93	0.69	0.79
2	0.95	0.97	0.96
3	0.62	0.72	0.67
4	0.55	0.74	0.63

Accuracy			**0.75**
Macro average	**0.78**	**0.75**	**0.76**
Weighted average	**0.78**	**0.75**	**0.76**

Table 6 shows the performance for the CNB classifier demonstrating a large variability of performance in the classes of bill and mandatory health plan with 30 % and 31 %, respectively, depending on the analysis on the metrics of Precision and Recall. In contrast, the health system in Colombia class is the most consistent class across metrics.

Table 6: CNB classifier performance

Class	Precision	Recall	F1-score
0	0.93	0.62	0.74
1	1.00	0.72	0.84
2	0.85	1.00	0.92
3	0.69	0.61	0.65
4	0.52	0.82	0.63
Accuracy			**0.75**
Macro average	**0.8**	**0.75**	**0.76**
Weighted average	**0.8**	**0.75**	**0.76**

Table 7 shows the performance for the RF classifier, demonstrating a small range depending on the class to be analyzed in the documents correctly classified. In both metrics, medicine documents obtained the same percentage, placing it as the best class in Precision and second after tutela actions if analyzed under the Recall metric.

Table 7: RF classifier performance

Class	Precision	Recall	F1-score
0	0.96	0.57	0.72
1	0.62	1.00	0.76
2	0.95	0.95	0.95
3	0.89	0.67	0.76
4	0.80	0.84	0.82
Accuracy			**0.80**
Macro average	**0.84**	**0.81**	**0.80**
Weighted average	**0.84**	**0.80**	**0.80**

Table 8 shows the performance for the SVM classifier, demonstrating very similar levels of correctly classified documents in the classes of tutela action, the health system in Colombia, and the mandatory health plan. As with the LR classifier, the tutela action and medicine documents are the best performing classes in their respective performance metrics. When looking at the Recall metric on correctly identified documents, the lowest levels can be observed.

Table 8: SVM classifier performance

Class	Precision	Recall	F1-score
0	0.81	0.62	0.70
1	0.72	1.00	0.84
2	1.00	0.95	0.97
3	0.75	0.67	0.71
4	0.74	0.76	0.75
Accuracy			**0.80**
Macro average	**0.81**	**0.80**	**0.79**
Weighted average	**0.81**	**0.80**	**0.80**

Table 9 shows the ideal metrics for comparing the classification models, Accuracy, and Weighted average F1-score. Both metrics indicate the overall performance of the classifier; however, the use of the metrics depends on the researcher's approach. Accuracy focuses on the so-called true positives and true negatives; that is, those documents that were classified under the correct class and documents that were not classified in the class of analysis because they were not that type of document. In this article, F1-score is the appropriate metric to perform the interpretations because the models do not balance the samples across classes. Table 9 shows the best classification performance in RL followed by a tie between the random RF and SVM models, showing a 4% difference. Similar to the second place, the third place is occupied by a tie between NBC and CBN.

Table 9: Comparing classification models

Model/ Metric	LR	NBM	CBN	RF	SVM
Accuracy	**0.85**	0.75	0.75	0.80	0.80
Weighted average F1-score	**0.84**	0.76	0.76	0.80	0.80

In the following paragraphs, there is a discussion in the light of the cited papers shown in table 10 that performed the main task, the classification of documents. In addition, paragraphs contain explanations of the literature conflicts to support some evaluations provided in the discussion. Among

the literature review section, the articles binding to the discussion are mostly found within the descriptive subsections of the relation of their contributions to the main task, i.e., subsections: Text mining and document, Information retrieval and document classification, and Text summarization and document classifications are responsible subsections for contributing such binding articles.

Table 10. Performance comparison of cited papers across different approaches.

Author	Approach	Results
Dewi, K. and Sagala, R (2018)	Fuzzy K-Nearest Neighbor Classification	25% by summary; 20% by without summary using both 10 k-levels
Khalifi et al. (2018)	Combined k-means and Support vector machines (SVM)	100% approach with; 97.75% approach without
Kumar et al. (2017)	Adaptive Neuro-Fuzzy Inference System (ANFIS)	68.52% approach with; 56.10% Neural Network; 51.48% Fuzzy Logic
Liu et al. (2007)	Sense-based text classification algorithm	32%
Mamakis et al. (2012)	A supervised classifier that assigning a random document to a class with Naive Bayes Assumption of statistical independences	92.35% approach with; 85.55% NBC; 80.45% LM-6; 83.29% LM-3
Tang et al. (2015)	User Product Neural Network (UPNN)	43.5% for IMDB Dataset; 60.8% Yelp2014 Dataset; 59.6% Yelp 2013 Dataset
Thanh et al. (2015)	Semi-supervised learning Super Vector Machine (SVM)	96.22% with approach; 95.96% L2-SVM; 88.23% RLS
Wen et al. (2020)	Speculative sentiment classification model	57.8% IMDB; 68.1% Yelp2013; 69.0% Yelp2014
Xu et al. (2016)	Cached Long Short-Term Memory Neural Networks	59.8% Yelp2013; 60.9% Yellp2014; 44.9% IMDB
Yang et al. (2016)	Hierarchical attention network	68.2% Yelp2013,70.5% Yelp2014, 71.0% Yelp2015; 49.4% IMDB; 75.8% Yahoo answers, 63.6 Amazon
Yang et al. (2020)	Semantic document classification based on SSC to improve ML algorithms	79.51% Rotten Tomatoes dataset; 83.6% THUCNews

The article's performance is ranked fourth in ranking the best models compared to the cited articles. The outstanding performances of the first three models contain databases with a short length in their documents. Khalifi et al. [20] ran their supervised model with a Yahoo corpus, Thanh et al. [28] fit a semi-supervised model with news from a Vietnamese newspaper (vnexpress.net), and Mamakis et al. [26] applied their normalization approach to a database to online news from a Greek newspaper. Thus, the model performance is preferred when one wants to implement a classifier on longer documents.

In contrast, similar cases in length have much more lagging performances than Dewi et al. [29], with a dataset of 100 theses presenting a maximum performance of 25% for a ten k-level, and Liu et al. perform their sense-based algorithm with a labeled Brown corpus. Finally, Altinel and Ganiz [33] perform a survey on the implementation of different models focusing on document classification; therefore, table 11 shows some articles that also used the social network Wikipedia as ontology by way of improving existing classifiers.

Table 11 Recompiled articles with Wikipedia Ontology from Atinel and Ganiz [33]

Author	Approach	Results
Gabrilovich and Markovitch (2007)	Análisis Semantico Explicito	72% Autralian Broadcasting Corporation's news mail service
Wang and Domeniconi (2008)	Supervised (Multi-words with strategy and Linear Kernel)	89.92% 20NewsGroup
Yang et al. (2013)	Supervised SVM	93% Ohsumed collection which includes medical abstracts from MeSH
Sungaya and Gomati (2013)	Supervised Multilayer SVM +kNN-NB	93.0% 20NewsGroup
torunoğlu et al. (2013)	NB	74% Twitter Sentiment 140 dataset

As in the case of the cited studies, the articles that used Wikipedia as the ontology presenting the best performances are based on short datasets in the news except for Yang et al. [38]. However, the methodology approach used in these articles in table 11 comprises modifications to the classifiers used for this article.

5. CONCLUSIONS

In this paper, we address a system that solves the document classification problem for the Spanish language by building the classes with public sources by tracking Google and Wikipedia searches for binding documents to Colombian health sector. The empirical results showed outstanding performance with a small range of variability among the different classification models performed.

Very few studies have as a source of information a public social network such as Wikipedia to build a database adding a contribution to the analysis of documents in the Spanish language; instead, Wikipedia used to be an ontology for the improvement of existing ML classifiers. In addition, the current knowledge presents several limitations; the first one is the scarcity of application of this source of information (Wikipedia) to perform classifications in large documents; secondly, the number of documents explicitly collected to Colombia does not equal the number of other databases present in the literature review conducted to improve the performance of the classifiers presented in this article.

For future work, we will compare the performances of the set of models performed with a balanced dataset vs. deep learning models to demonstrate the benefits and weaknesses of these models. In addition, the results of the classification system can be input for the development of one of the related tasks: TM, IR, or TS.

ACKNOWLEDGEMENT

The paper was produced thanks to the joint work of the International Research Group in Computer Science, Communications and Knowledge Management (GICOGE) of the Universidad Distrital Francisco José de Caldas, LUMON LV TECH research group, the SUOMAYA CGMLTI SENA research group.

REFERENCES

[1] J. Han, M. Kamber, and J. Pei, "Data Mining Trends and Research Frontiers," in *Data Mining*, 2012, pp. 585–631.

[2] M. R. M. Talabis, R. McPherson, I. Miyamoto, J. L. Martin, and D. Kaye, "Analytics Defined," in *Information Security Analytics*, 2015, pp. 1–12.

[3] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey, "Scatter / Gather: A Cluster-based Approach to Browsing Large Document Collections," *Spec. Interes. Gr. Inf. Retr.*, vol. 51, no. 2, pp. 318–329, 1992, [Online]. Available: http://doi.acm.org/10.1145/133160.133214.

[4] M. A. Hearst, "Untangling text data mining," in *ACL '99: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, 1999, pp. 3–10, doi: 10.3115/1034678.1034679.

[5] V. Kotu and B. Deshpande, "Text mining," in *Data Science*, 2019, pp. 281–305.

[6] S. L. Ting, W. H. Ip, and A. H. C. Tsang, "Is Naïve bayes a good classifier for document

classification?," *Int. J. Softw. Eng. its Appl.*, vol. 5, no. 3, pp. 37–46, 2011.

[7] I. H. Witten *et al.*, "Moving on: applications and beyond," *Data Min.*, pp. 503–532, 2017, doi: 10.1016/B978-0-12-804291-5.00013-1.

[8] D. Tang, B. Qin, and T. Liu, "Learning semantic representations of users and products for document level sentiment classification," in *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*, 2015, vol. 1, pp. 1014–1023, doi: 10.3115/v1/p15-1098.

[9] J. Xu, D. Chen, X. Qiu, and X. Huang, "Cached long short-term memory neural networks for document-level sentiment classification," in *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2016, pp. 1660–1669, doi: 10.18653/v1/d16-1172.

[10] J. Wen, G. Zhang, H. Zhang, W. Yin, and J. Ma, "Speculative text mining for document-level sentiment classification," *Neurocomputing*, vol. 412, pp. 52–62, 2020, doi: 10.1016/j.neucom.2020.06.024.

[11] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical Attention Networks for Document Classification," in *Proceedings ofNAACL-HLT 2016*, 2016, pp. 1480–1489, [Online]. Available: http://arxiv.org/abs/1606.02393.

[12] J. Liu *et al.*, "Data Mining and Information Retrieval in the 21st century: A bibliographic review," *Comput. Sci. Rev.*, vol. 34, pp. 1–13, 2019, doi: 10.1016/j.cosrev.2019.100193.

[13] K. H. Ambert and A. M. Cohen, "Text-Mining and Neuroscience," in *International Review of Neurobiology*, 1st ed., vol. 103, Elsevier Inc., 2012, pp. 109–132.

[14] J. Atkinson-Abutridy, C. Mellish, and S. Aitken, "A semantically guided and domain-independent evolutionary model for knowledge discovery from texts," *IEEE Trans. Evol. Comput.*, vol. 7, no. 6, pp. 546–560, 2003, doi: 10.1109/TEVC.2003.819262.

[15] M. Pérez-Montoro and L. Codina, "The Essentials of Search Engine Optimization," in *Navigation Design and SEO for Content-Intensive Websites*, 2017, pp. 109–124.

[16] C. Tenopir, P. Wang, Y. Zhang, B. Simmons, and R. Pollard, "Academic users' interactions with ScienceDirect in search tasks: Affective and cognitive behaviors," *Inf. Process. Manag.*, vol. 44, no. 1, pp. 105–121, 2008, doi: 10.1016/j.ipm.2006.10.007.

[17] Y. Zhang, "The development of users' mental models of MedlinePlus in information searching," *Libr. Inf. Sci. Res.*, vol. 35, no. 2, pp. 159–170, 2013, doi: 10.1016/j.lisr.2012.11.004.

[18] A. Gordo, F. Perronnin, and E. Valveny, "Large-scale document image retrieval and classification with runlength histograms and binary embeddings," *Pattern Recognit.*, vol. 46, no. 7, pp. 1898–1905, 2013, doi: 10.1016/j.patcog.2012.12.004.

[19] S. S. Lin, "A document classification and retrieval system for R&D in semiconductor industry - A hybrid approach," *Expert Syst. Appl.*, vol. 36, pp. 4753–4764, 2009, doi: 10.1016/j.eswa.2008.06.024.

[20] H. Khalifi, A. Elqadi, and Y. Ghanou, "Support vector machines for a new hybrid information retrieval system," in *The First International Conference On Intelligent Computing in Data Sciences Support*, 2018, vol. 127, pp. 139–145, doi: 10.1016/j.procs.2018.01.108.

[21] V. N. Gudivada, "Natural Language Core Tasks and Applications," in *Handbook of Statistics*, 1st ed., vol. 38, Elsevier B.V., 2018, pp. 403–428.

[22] W. Song, L. Cheon Choi, S. Cheol Park, and X. Feng Ding, "Fuzzy evolutionary optimization modeling and its applications to unsupervised categorization and extractive summarization," *Expert Syst. Appl.*, vol. 38, no. 8, pp. 9112–9121, 2011, doi: 10.1016/j.eswa.2010.12.102.

[23] H. Zhuge, *Multi-Dimensional Summarization in Cyber-Physical Society*. Elsevier, 2016.

[24] P. D. Karp, "An ontology for biological function based on molecular interactions," *Bioinforma. Ontol.*, vol. 16, no. 3, pp. 269–285, 2000, doi: 10.1093/bioinformatics/16.3.269.

[25] S. Theodoridis, "Learning in Reproducing Kernel Hilbert Spaces," in *Machine Learning*, 2020, pp. 531–594.

[26] G. Mamakis, A. G. Malamos, J. A. Ware, and I. Karelli, "Document Classification in Summarization," *J. Inf. Comput. Sci.*, vol. 7,

no. 1, pp. 25–36, 2012.

[27] Y. J. Kumar, F. J. Kang, O. S. Goh, and A. Khan, "Text Summarization Based on Classification Using ANFIS," *Stud. Comput. Intell.*, vol. 710, pp. 405–417, 2017, doi: 10.1007/978-3-319-56660-3_35.

[28] H. K. H. Vo Duy Thanh, Vo Trung Hung and Tran Quoc Huy, "Text Classification Based on SVM and Text Summarization," *Int. J. Eng. Res. Technol.*, vol. 4, no. 2, pp. 181–186, 2015, [Online]. Available: www.ijert.org.

[29] K. E. Dewi and R. E. Sagala, "Using Summarization to Optimize Text Classification," in *IOP Conference Series: Materials Science and Engineering*, 2018, vol. 407, no. 1, p. 6, doi: 10.1088/1757-899X/407/1/012157.

[30] I. H. Witten and E. Frank, "Algorithms: the basic methods," in *Data Mining: practical Machine Learning Tools and Techniques*, Second Edi., 2005, pp. 97–105.

[31] A. Khan, B. Baharudin, L. H. Lee, and K. Khan, "A Review of Machine Learning Algorithms for Text-Documents Classification," *J. Adv. Inf. Technol.*, vol. 1, no. 1, pp. 4–20, 2010, doi: 10.4304/jait.1.1.4-20.

[32] S. Yang, R. Wei, J. Guo, and H. Tan, "Chinese semantic document classification based on strategies of semantic similarity computation and correlation analysis," *J. Web Semant. Sci. Serv. Agents World Wide Web*, vol. 63, pp. 1–15, 2020, doi: 10.1016/j.websem.2020.100578.

[33] B. Altınel and M. C. Ganiz, "Semantic text classification: A survey of past and recent advances," *Inf. Process. Manag.*, vol. 54, no. 6, pp. 1129–1153, 2018, doi: 10.1016/j.ipm.2018.08.001.

[34] Y. Liu, P. Scheuermann, X. Li, and X. Zhu, "Using WordNet to disambiguate word senses for text classification," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4489 LNCS, no. PART 3, pp. 781–789, 2007, doi: 10.1007/978-3-540-72588-6_127.

[35] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Inf.*, vol. 10, no. 4, pp. 1–68, 2019, doi: 10.3390/info10040150.

[36] L. Xu, "Case study: Image clustering," in *Heterogeneous Computing with OpenCL 2.0: Third Edition*, D. Kaeli, P. Mistry, D. Schaa, and D. Zhang, Eds. Morgan Kaufmann Publishers, 2015, pp. 213–228.

[37] R. Guerrero, A. I. Gallego, V. Becerril-Montekio, and J. Vásquez, "Sistema de salud de Colombia," *Salud Publica Mex.*, vol. 53, no. SUPPL. 2, pp. 1–12, 2011, doi: 10.1590/s0036-36342011000500003.

[38] Yang, L., Li, C., Ding, Q., Li, L.: Combining lexical and semantic features for short text classification. ProcediaComputer Science22(Complete), 78–86 (2013). doi:10.1016/j.procs.2013.09.083