NOVA IMS Information Management School

MAA

Mestrado em Métodos Analíticos Avançados
Master Program in Advanced Analytics

PORTUGUESE PATENT CLASSIFICATION

A use case of text classification using machine learning and transfer learning approaches

Ádria Lidiane de Oliveira Alves Ferreira

Project Work presented as partial requirement for obtaining the Master's degree in Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

PORTUGUESE PATENT CLASSIFICATION

A USE CASE OF TEXT CLASSIFICATION USING MACHINE LEARNING AND TRANSFER LEARNING APPROACHES

by

Ádria Lidiane de Oliveira Alves Ferreira

Project Work presented as a partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics

Advisor: Roberto André Pereira Henriques

Junho 2021

ACKNOWLEDGEMENTS

I would like to express my gratitude to my advisor Dr Roberto Henriques for all of the guidance and useful inputs during this thesis. I am also grateful for the chance to study at the Nova University Lisbon, and particularly for all the professors, staff, and colleagues who somehow contributed to this journey.

My deep thanks to Jorge Rodrigues da Ponte, former INPI Board Member, and Inês Cristóvão da Silva, a member of the INPI team, for all their availability and valuable knowledge. To André Marques and my colleagues for their support, suggestions, and friendly talks in these tough months of 2020/2021.

I would like to thank my friends, from there and from here, for all their encouragement and care during this time.

Special thanks to my family for always being by my side, even with an ocean away.

To God. "For from him, and through him, and to him, are all things" (Rm 11:36a)

ABSTRACT

Patent classification is one of the areas in Intellectual Property Analytics (IPA), and a growing use case since the number of patent applications has been increasing through the years worldwide. Patents are more than ever being used as financial protection for companies that also use patent databases to raise researches and leverage product innovations. Instituto Nacional de Propriedade Industrial, INPI, is the government agency responsible for protecting Industrial Property rights in Portugal. INPI has promoted a competition to explore technologies to solve some challenges related to Industrial Properties, including the classification of patents, one of the critical phases of the grant patent process.

In this work project, we used the dataset put available by INPI to explore traditional machine learning algorithms to classify Portuguese patents and evaluate the performance of transfer learning methodologies to solve this task. BERTTimbau, a BERT architecture model pre-trained on a large Portuguese corpus, presented the best results to the task, even though with a performance only 4% superior to a LinearSVC model using TF-IDF feature engineering. In general, the model presents a good performance, despite the low score when classes had few training samples. However, the analysis of misclassified samples showed that the specificity of the context has more influence on the learning than the number of samples itself.

Patent classification is a challenging task not just because of 1) the hierarchical structure of the classification but also because of 2) the way a patent is described, 3) the overlap of the contexts, and 4) the underrepresentation of the classes. Nevertheless, it is an area of growing interest, and that can be leveraged by the new researches that are revolutionizing machine learning applications, especially text mining.

KEYWORDS

Natural Language Processing (NLP); Text Mining; Patent classification; Transfer Learning; Bi-directional Encoder Representations for Transformers (BERT)

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ATN	Augmented Transition Network
BERT	Bi-directional Encoder Representations for Transformers
BoW	Bag-of-Words
BrWaC	Brazilian Web as a Corpus
CPC	Cooperative Patent Classification
EPO	European Patent Office
GLUE	General Language Understanding Evaluation
INPI	Instituto Nacional de Propriedade Industrial
IPA	Intellectual Property Analytics
IPC	International Patent Classification
LM	Language Model
LSTM	Long Short-Term Memory
MLM	Masked Language Modeling
NLM	Neural Language Model
NLP	Natural Language Processing
NSP	Next Sentence Prediction
SLM	Statistical Language Model
SQuAD	Stanford Question Answering Dataset
TF-IDF	Term Frequency-Inverse document frequency
ULMFiT	Universal Language Model Fine-tuning
USPTO	United States Patent and Trademark Office
WIPO	World Intellectual Property Organization

1. INTRODUCTION

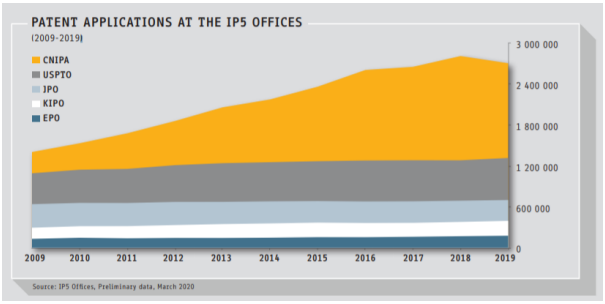
Intellectual Property (IP) is a category of property related to the "creations of the mind", it means a vast range of activities from art to scientific works, trademarks, and inventions. It aims to promote the development of intellectual goods while giving the creators economic rights over their creations for a certain period. IP is divided into two main types, i) copyright and related types, which covers scientific, artistic, and literary works; ii) industrial property, which includes patents, trademarks, industrial designs, and geographical indications (Wipo, [s.d.]).

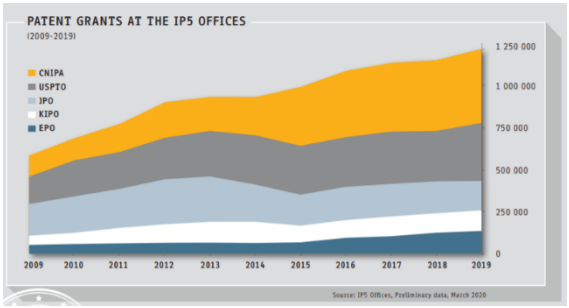
Intellectual Property Analytics (IPA) is a growing field that deals with the analysis of intellectual property databases to discover trends, relationships and patterns, and to leverage researches and innovations based on information that may not be available anywhere else (Aristodemou e Tietze, 2018).

A patent is an intellectual property right granted for the protection of an invention. According to WIPO[1], "an invention can be a product or a process that provides, in general, a new way of doing something, or offers a new technical solution to a problem." Once a patent is registered, the holder has the exclusive right to produce and commercialize it for a certain period. Therefore it becomes financial protection for companies that also use it as a strategic resource for information and knowledge management (Trappey, Trappey, Wu e Lin, 2012). Patent databases can be used to raise research and development activities, product innovations, or technology transfer (Li, Hu, Cui e Hu, 2018).

A patent is a territorial right; hence the process to obtain it is regulated by each country and conducted by a specific patent office. In general, to gain this protection, the inventor will need to file an application describing the invention and submit it to the analysis of utility, novelty, and inventiveness (Wipo, [s.d.]).

The growth of applications through the years turns the patent analysis into one of the areas of IPA with significant relevance. In its preliminary statistic report from 2019, IP5[2] reported that 2.7 million patent applications were filed at its offices, and 1.25 million were granted, an increase of 6% compared to the previous year. (Figure 1.1)





[1] WIPO – World Intellectual Property Organization: the global forum for intellectual property services, policy, information and cooperation. A self-funding agency of the United Nations, with 193 member states (https://www.wipo.int/)

[2] IP5 - The five IP offices: is the name given to a forum of the five largest intellectual property offices in the world that was set up to improve the efficiency of the examination process for patents worldwide

Figure 1.1 - Evolution of patent applications and grants at IP5 offices from 2009 to 2019 (*IP5 Statistics Report*, 2019)

EPO [3]received a record number of applications in 2019, with a yearly growth of 4%. Half of all patent applications came from companies based in Europe, and among those, more than 70% were from large companies. (Figure 1.2) The strongest growth came from the Digital communication field, followed by computer technology, driven by AI. (Figure 1.3)

Figure 1.2 - Patent application distribution by type of applicant (*EPO - Patent Index 2019*, 2019)

The patent application is a lengthy document with all the details about the invention. It contains the definition of the invention and its delimitations, structured in several sections, namely title, abstract, description, and claims. While the abstract gives a summary about the patent, the description brings enough details that allow one to use the invention after the expiration of the patent's term. The claims, on the other hand, define the scope of a patent and help to determine the extent of protection to be granted by the patent (*Código de Propriedade Intelectual*, 2019).

Figure 1.3 - EPO distribution of granted patents in 2019 by field (*EPO - Patent Index 2019*, 2019)

The process to obtain a patent granted takes several months and starts with its classification. The patent classification is a formal examination of the application document conducted by an expert who sets one or more categories to the patent request based on its content. This first step consumes

[3] EPO - European Patent Office: is the executive arm of the European Patent Organisation, an international intergovernmental organisation with 38-member states.

a big part of the processing time and demands vast knowledge of the classification system. Meanwhile, it is crucial in the following analysis of the invention's originality (*Código de Propriedade Intelectual*, 2019). The classification system is also fundamental for patent analysis by helping the navigation among the patents and enabling to perform more accurate searches. Furthermore, it facilitates retrieval tasks across different languages and patent offices since it is a multi-language code (Gomez e Moens, 2014).

1.1. PATENT CLASSIFICATION SYSTEM

Administered by WIPO, International Patent Classification (IPC) is the most important classification system and so the primary reference for patent classification. The areas of technology are split across eight sections (Table 1.1) and each section is subdivided into classes, subclasses, groups, and subgroups, counting in the last level approximately 72,000 sub-groups (*Espacenet - Home page*, [s.d.]). An example of IPC hierarchical structure can be seen in Figure 1.4.

Section	Description
A	Human Necessities
B	Performing Operations; Transporting
C	Chemistry; Metallurgy
D	Textiles; Paper
E	Fixed Constructions
F	Mechanical Engineering; Lighting; Heating; Weapons; Blasting Engines or Pumps
G	Physics
H	Electricity

Table 1.1 - IPC Areas of Technology

Cooperative Patent Classification (CPC) system is an initiative of EPO and USPTO[4] that took IPC as a basis "to harmonize their classification systems with a similar structure to the IPC but more detailed than it to improve the patent searching". A section 'Y' was added to CPC for emerging technology and cross-sectional technologies spanning over several sections of the IPC (*Cooperative Patent Classification*, [s.d.]). CPC has been adopted also by other patent offices and one believes that eventually, it will replace the IPC system (Lee e Hsiang, 2020).

Patent classification is challenging not just because of the hierarchical structure, and the high number of sub-areas in the last level but also because one patent can be related to more than one subclass, group, or subgroup, which makes it a multi-label classification task. Moreover, the

[4] United States Patent and Trademark Office - USPTO: is the federal agency for granting U.S. patents and registering trademarks.

distribution among the categories is highly unbalanced and tends to follow a Pareto-like distribution, where about 80% of the documents fall into about 20% of the categories. In addition, a patent application is a document with legal and technical terminologies and specific writing style, on which rare words help to give the idea of novelty and avoid plagiarism (Abdelgawad, Kluegl, Genc, Falkner e Hutter, 2020; Li, Hu, Cui e Hu, 2018).

Structure	Symbol	Description
Section	**A**	Human necessities
Class	**A21**	Baking; edible doughs
Subclass	**A21C**	Machines or equipment for processing doughs
Group	**A21C1**	Mixing or kneading machine for the preparation of dough
Subgroup	**A21C1/06**	With horizontally-mounted mixing or kneading tools

Figure 1.4 - Example of IPC hierarchical structure – A21C1/06

1.2. PATENTS IN PORTUGAL

In Portugal, Instituto Nacional de Propriedade Industrial, INPI[5], is the government agency responsible for protecting Industrial Property rights. The institute was created in 1976 and is responsible for granting exclusive rights not just on patents but also on trademarks and designs. It also represents the country in international organizations and contributes to the modernization of the business community, promoting innovation and competitiveness.

Annually, INPI discloses statistical reports of Industrial Property in Portugal. After some decrease at the beginning of the decade, it has been reported an increase of patents applications since 2016, with a growth of 40% from 2019 compared to the previous year. (Figure 1.5)

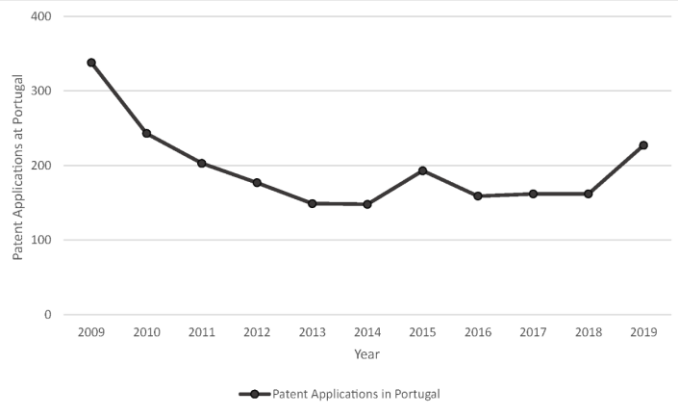


Figure 1.5 - Patents applications in Portugal since 2009

[5] https://inpi.justica.gov.pt/

In 2019, just like the EPO's trend, 40% of the patent applications came from companies, followed by independent inventors (35%) and Higher Education Institutions (18.4%). Most of the patent applications are in Human Necessities (sector A), on special because this sector covers a wide field of utilization. In addition to sector A, Chemistry/Metallurgy (sector C) and Transporting/Performing Operations (sector B) represent almost 80% of the requests, with similar behavior in the previous years. (Figure 1.6)

The process to obtain a patent granted takes at least 21 months and starts with a formal examination conducted by an expert who applies one or more categories to the patent request based on keywords extracted from the claims section. In the classification task, INPI experts use local tools as well as intelligent tools provided by WIPO.

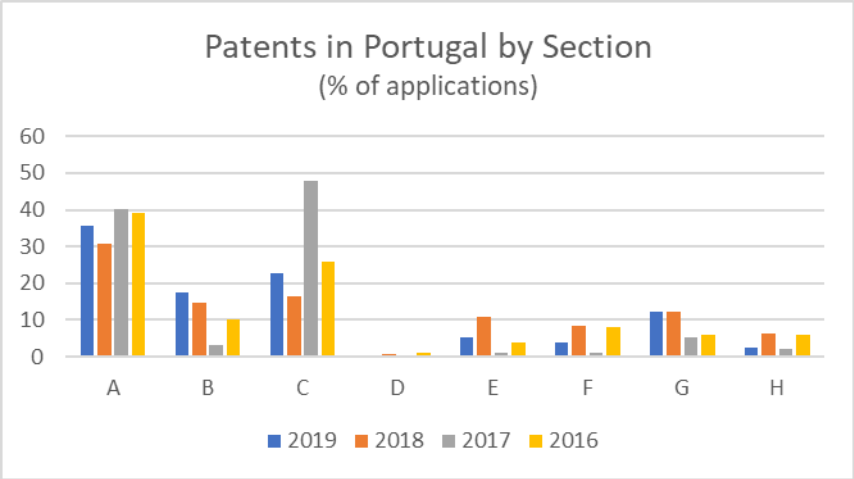


Figure 1.6 - Portuguese granted patents by IPC section

It is a critical phase because the applied classification will guide the next phase when the invention proposal is analysed and compared to existing patents to guarantee that it is not yet protected. Then, the application is published in the Industrial Property Bulletin, and interested parties have the opportunity to oppose it. After that, a broad examination is accomplished, and the decision is taken.

The agency has been exploring intelligent tools to increase the efficiency of its tasks. Recently, INPI promoted a contest to explore new methods to classify patents. A data set of granted pre-classified patents were made available to the participants, and the goal was to present a solution to classify patents in the second level of IPC.

1.3. Study Objectives

According to Feldman and Sanger, the process of discovering information in extensive text collections to identify relationships and patterns automatically can be understood as text mining. This interdisciplinary area uses Natural Language Processing (NLP) and machine learning or deep learning algorithms to explore data sources and extract or create information from them (Feldman e Sanger, 2006).

However, as the data sources are unstructured or semi-structured text data from document collections, the preprocessing phase that deals with the transformations of text into structured data sets becomes a crucial aspect of it. The success of text mining is highly dependent on these operations. The challenge here is how to represent the complex characteristics of word uses and their variations across linguistics contexts in a structured way. It is not a trivial task since language by itself is complex, and its production and comprehension traverse many levels of linguistic analysis, like morphological, lexical, syntactical, semantical, and pragmatical. (Feldman e Sanger, 2006; Liddy, 2001; Peters, Neumann, Iyyer, Gardner, Clark, Lee e Zettlemoyer, 2018).

The amount of textual information available for centuries summed to the increase of information created or loaded electronically has been boosting the advances in text mining (Feldman e Sanger, 2006; Korde, 2012). Statistical methods and traditional machine learning algorithms can be a good approach in some scenarios with specific contexts. Moreover, researches with deep learning, high-quality language models and word vectors have led to substantial gains in this field (Manning, 2020; Mikolov, Chen, Corrado e Dean, 2013).

Nevertheless, the lack of labelled training instances or the cost to train a model with a massive amount of data can be a limitation for specific real-world applications. To overcome these challenges, the strategy of transferring the knowledge across domains instead of training from scratch, which had a large impact on computer vision, has been revolutionizing text mining. The transfer learning approach and pre-trained language models have been achieving the state-of-art in different tasks (Pan e Yang, 2010; Zhuang, Qi, Duan, Xi, Zhu, Zhu, Xiong e He, 2021).

This work project aims to solve the problem of classifying Portuguese patents proposed by INPI using text mining. In this study, the title, abstract, and claims of Portuguese patents are used to train a model and predict their class on the second level of the International Patent Classification system (IPC).

Different approaches to represent the normalized text as vectors that will be the input of the classification algorithms will be applied and assessed. Then, machine learning and deep learning classification algorithms will be explored and their performances compared. Finally, pre-trained language models will be tuned on this dataset to evaluate the transfer learning approach in this real-case scenario.

Objectives:

- Explore a dataset of granted patents and prepare it to be used in classification models.
- Evaluate machine learning and deep learning algorithms to solve the problem.
- Using transfer learning methodologies, fine-tune pre-trained language models for automatically classifying Portuguese patents.
- Assess the best model and analyze the results to identify the most relevant attributes of the classification process.

This document is organized into six sections. The first chapter introduced the main topic with an overview of the context in which the case study was developed; brought an explanation about the

patent, its classification system, and the patent granted evolution; and presented the work project objectives. The second chapter includes theoretical concepts on which the work project is based and related studies about patent classification. Chapter three brings more details of the study case, the methodology, techniques and models applied. The results are demonstrated and discussed in chapter four. In the fifth chapter, the conclusions are presented and then, future work is proposed in the sixth chapter.

2. LITERATURE REVIEW

In this chapter, we will present the concepts of Natural Language Processing (NLP), text classification, language models and explore the uses of transfer learning on NLP with an overview of a specific pre-trained language model, BERT (Bidirectional Encoder Representations from Transformers). Also, related studies about patent classification are presented.

2.1. NATURAL LANGUAGE PROCESSING

For Khurana et al., "Natural Language Processing (NLP) is a tract of Artificial Intelligence and Linguistics devoted to making computers understand the statements or words written in human language"(Khurana, Koli, Khatter e Singh, 2017). Researches in NLP date back to 1940, initially focused on machine translation systems. In the late 1960s, after some period of disrepute, some significant developments, both in theoretical issues and in the construction of prototype systems, were done. Next, semantic gained attention, and by the end of the 1980s, statistical approaches were shown to be complementary in many respects to symbolic approaches (Khurana, Koli, Khatter e Singh, 2017; Liddy, 2001; Nadkarni, Ohno-machado e Chapman, 2011). According to Liddy, while in symbolic approaches, the analysis of linguistic phenomena is based "on explicit representation of facts about language through well-understood knowledge representation schemes and associated algorithms", like in logic or rule-based systems; statistical approaches develop "approximate generalized models of linguistic phenomena based on actual examples of these phenomena provided by the text corpora without adding significant linguistic or world knowledge", using mathematical techniques and large text corpora (Liddy, 2001).

NLP can be applied in a diverse range of tasks, from translation to sentiment analysis, each one with its challenges. Some frequent applications are:

- Machine translation: refers to the automatic translation of text from one human language to another. The challenge here is to keep the meaning of sentences intact along with grammar and tenses (Khurana, Koli, Khatter e Singh, 2017).

- Named Entity Recognition: recognition, tagging, and extraction into a structured representation, certain key elements of information, e.g. persons, companies, locations, organizations. The understanding of context to differ the fruit "apple" for the brand "Apple", or the analysis of n-grams to identify "New York" instead of "new" and "York", are the biggest challenges in this task (Liddy, 2001).

- Summarization: reduces a large text, extracting its key phrases, yet keeping the main meaning (Liddy, 2001).

- Sentiment Analysis: focuses on identifying sentiments about a given topic. The challenge in this task is to identify the sentiment even when it is not explicit or when the text contains irony or sarcasm. Sometimes a hard task even for humans (Khurana, Koli, Khatter e Singh, 2017).

- Text Classification: categorize a document in a predefined set of categories or classes. Capturing the whole context can not be so simple in this task (Silva e Ribeiro, 2010).

2.1.1. Text Classification

Text classification has been one of the most known text mining downstream tasks. Initially, text classification was used to be solved manually, based on hand-coded rules, created and maintained by a domain expert, which made it labour-intense, sometimes high costly and demanding. Intelligent text classification methods, though, provide superior facilities, save time and money while handling the increase of digital texts we are facing (Manning, Raghavan e Schutze, 2008; Silva e Ribeiro, 2010).

As a supervised classification task, the training set consists of text data, and each document is labeled with a class value from a set of discrete values. It may be formalized as the task of approximating the unknown target function $f: D * C \rightarrow \{-1,1\}$ where f is the text classifier, $C = \{c1, c2, \ldots, cn\}$ is a predefined set of categories or classes and D is a set of documents. It can be a binary or multiclass problem and either a single-label or multi-label task (Silva e Ribeiro, 2010).

Text classification solutions typically follow four phases: Feature engineering, dimensions reductions, classifier selection and evaluation. The first step is to create a structured set for our training purposes. This is a crucial activity that involves cleaning and preprocessing the text and selecting the best form of representing the words and thus, each document. Dimensionality reduction is an optional step that can help to reduce the time and memory complexity while maintaining the main characteristics of the data (Kowsari, Meimandi, Heidarysafa, Mendu, Barnes e Brown, 2019).

The classifier selection is a critical phase and many classifiers have been used for text classification. Traditional Machine Learning algorithms, like Logistic Regression, Naïve Bayes Classifier, Support Vector Machine, k-nearest neighbor, have been applied with good results for many years. Ensemble-based learning techniques have also been successfully developed for document classification. And more recently, deep learning solutions, followed by transfer-learning approaches have been achieving state-of-the-art results. The evaluation phase allows us to understand and compare models' performance (Kowsari, Meimandi, Heidarysafa, Mendu, Barnes e Brown, 2019).

Some studies also add Data Acquisition, Data Analysis and Labelling as steps at the beginning of the text classification framework (Figure 2.1) (Mirończuk e Protasiewicz, 2018).

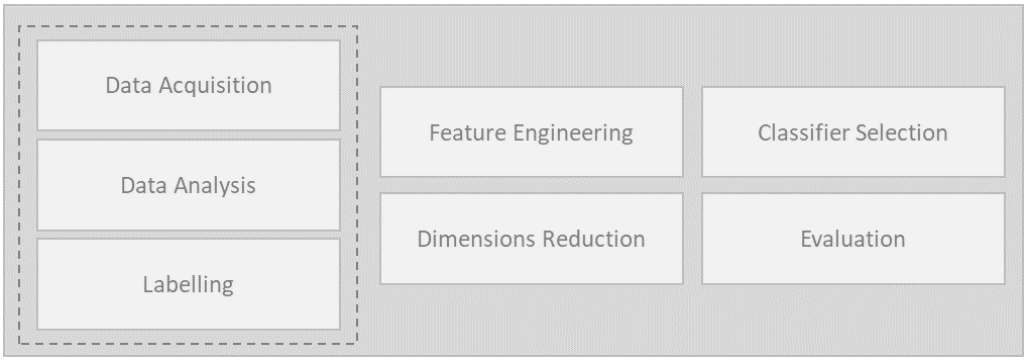


Figure 2.1 - Text classification phases

2.1.2. Language Models

Language Modeling is a task in NLP to predict the next token in a given sequence. A Language Model (LM) can be trained as a general-purpose feature extractor or trained in a large and general corpus to learn word representations (word embeddings) and be applied in different text mining tasks. A word embedding is a vector representation of a word in a high-dimensional vector space. The big advantage of embeddings is that one can compute degrees of similarity between two vectors and then compare documents in some collection (Merity, Keskar e Socher, 2018).

Statistical Language Model (SLM) uses a large amount of training data and employs statistical techniques to represent a linguistic unit. Methods such as bag-of-words (BoW) and N-grams models are well-known examples of SLM. In a term-document matrix, each document is represented as a count vector that can be built based on the frequency of the words in the corpus using methods like Term Frequency-Inverse Document Frequency (TF-IDF). On the other hand, the term-term matrix has the words correlated to other words that appear around it in a defined window in some training corpus. In both cases, the document and the word are represented as a sparse and long vector (Rosenfeld, 2000).

Their use of a large amount of text available electronically has resulted in simple but at the same time robust models that could outperform complex ones trained on fewer data. However, the categorical nature of language, the representation of the words as indices in a vocabulary with no notion of similarity and the high dimensionality of the representation can bring some limitations to it (Jurafsky e Martin, 2008; Korde, 2012; Rosenfeld, 2000).

For Howard and Ruder, an ideal Language Model should "capture many facets of language, such as long-term dependence, hierarchical relations and sentiment". Moreover, it should be "easily adapted to the idiosyncrasies of a target task". While in SLM the words are seen as atomic units, Neural Language Models (NLM) aim to contextualize them and then bring a better representation closer to the natural language representation. The approach, in this case, is to represent each word as a dense vector related to the surrounding words in such a way that the application of algebraic operations to capture words relationship is possible and useful (Howard e Ruder, 2018; Mikolov, Chen, Corrado e Dean, 2013).

Word2Vec was presented by Mikolov as a language model to compute "continuous vector representations of words from very large datasets". In the first proposed architecture, called the bag-of-words model (CBOW), the prediction of the current word is based on the context (words before and after), all words are projected into the same position, and the order of words does not influence the projection. The second proposed architecture, Skip-gram, aims to maximize a word classification based on another word in the same sentence. The current word is used as input in the projection layer, and the output is the probability from words with a certain range before and after the current one (Figure 2.3). To evaluate the models, different versions of word embeddings were compared, and several types of similarities tests were presented (Figure 2.3) (Mikolov, Chen, Corrado e Dean, 2013).

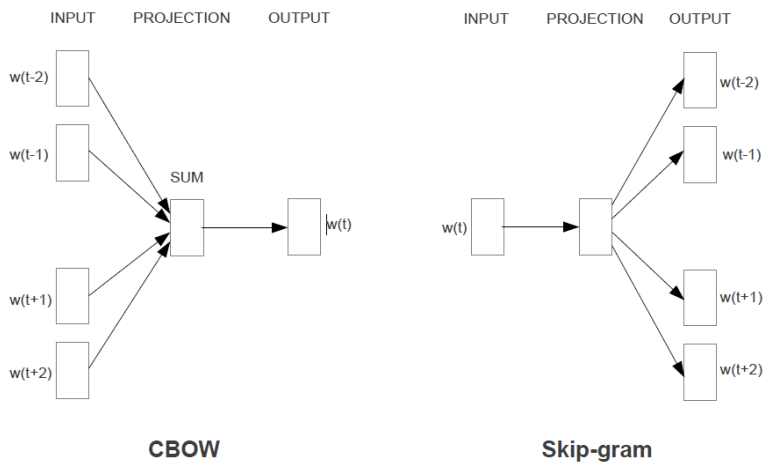


Figure 2.2 - Word2Vec architectures representation (Mikolov, Chen, Corrado e Dean, 2013)

Word2Vec boosted new discussions in the discipline, and after that, other essential embedding models were proposed, like GloVe – Global Vectors – that based on ratios of words co-occurrence probabilities combines the advantage of the global matrix factorization with skip-gram word analogy capture (Pennington, Socher e Manning, 2014), and FastText, proposed to overcome the limitation of models that do not handle with the morphology of words. In this model, each word is represented as a bag of character n-grams summed. Hence, this approach allows computing the embeddings for words that were not in the training data, a big advantage for morphologically rich languages (Bojanowski, Grave, Joulin e Mikolov, 2017).

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

Figure 2.3 - Examples of Word2Vec word pair relationships (Mikolov, Chen, Corrado e Dean, 2013)

Models like Glove and FastText use unsupervised learning techniques to learn relationships between words and to create embedding vectors. They are considered as context-window-based representations and can capture useful semantic and syntactic information. And, once one has pre-trained a model, the embeddings can be used in any task. Although, they missed an important aspect of linguistic, the polysemy – the possibility of multiple meanings of a word depending on the context (Peters, Neumann, Iyyer, Gardner, Clark, Lee e Zettlemoyer, 2018).

To address it, a deep contextualized word representation was introduced. ELMo (Embeddings from Language Model) is considered a semi-supervised model, trained in a bidirectional LSTM (Long Short-Term Memory), which representations are a function of this neural network internal layers for each

downstream task and computed using the entire input sentence and not just a range of words (Peters, Neumann, Iyyer, Gardner, Clark, Lee e Zettlemoyer, 2018).

Several deep learning architectures have been proposed and many researches have shown well-tuned neural networks to create LM using large-scale datasets and a vast amount of time and resources to achieve the state-of-art in different datasets from various languages. Besides, since it is an efficient method, another advantage of those models is the availability of pre-trained word embeddings from different languages, or even from specific contexts, ready to be used in any kind of text mining use case (Merity, Keskar e Socher, 2018).

2.2. TRANSFER LEARNING IN NLP

The idea of leveraging knowledge from one task and applying it to a different task to improve learning is the goal of transfer learning. Unlike human beings that can apply previous knowledge in new tasks, machine learning models usually learn each isolated task from scratch (Pan e Yang, 2010).

While Language Models can be seen as an important step in this direction and despite this approach has been widely used in computer vision, the applications of transfer learning in NLP had not shown a big impact and the models were still being trained from scratch with specific modification depend on the end task. The gain the ULMFiT (Universal Language Model Fine-tuning) brought was the introduction of "an effective transfer learning method" and fine-tuning techniques "that can be applied to any task in NLP" (Howard e Ruder, 2018). After ELMo provided a significant step towards pre-training, ULMFiT ushered the transfer learning/fine-tuning era on NLP.

ULMFiT did not present a new LM, instead, it used a known LM (AWD-LSTM) from (Merity, Keskar e Socher, 2018) with additional hyperparametrization, and pre-trained it in a large general-domain corpus. Then, as represented in Figure 2.4 this LM is fine-tuned on the data of the target task by the application of two methods proposed in the paper: discriminative fine-tuning and slanted triangular learning rates (STLR). The first one allows one to tune each layer with different learning rates. The second one is employed to make the model converge more quickly to the fitting parameter space region through the sequential linear increase and decay of the learning rate across the iterations. Finally, to perform the classification, two layers are added to the neural network and it is fine-tuned again using discriminative fine-tuning, STLR and gradual unfreezing. Those techniques aim to handle the most critical part of the transfer learning process: prevent catastrophic forgetting and slow conversion/overfitting, while enabling robust learning (Howard e Ruder, 2018).

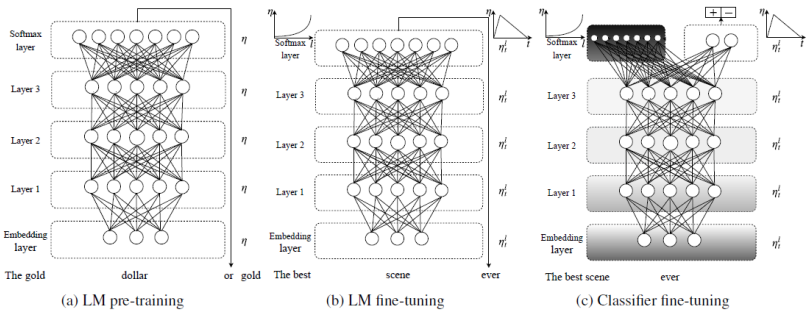


Figure 2.4 - ULMFiT stages representation (Howard e Ruder, 2018)

2.2.1. BERT

In the ULMFiT paper, the authors reveal the hope that those results "catalyze new developments in transfer learning for NLP" (Howard e Ruder, 2018). Indeed, some months later, researches scientists from Google AI Language, presented BERT (Bidirectional Encoder Representations from Transformers), that overcame state-of-the-art NLP systems performance from a variate of tasks, including on the Stanford Question Answering Dataset (SQuAD) and General Language Understanding Evaluation (GLUE) benchmark (*Google AI Blog: Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing*, [s.d.]).

Like ULMFiT, it works in two phases: pre-training, where the model is trained on a large-scale unlabeled dataset, and fine-tunning, when the downstream task dataset is used to refine the parameters. It uses completely different approaches for both phases, though, and an architecture based on Transformer (Devlin, Chang, Lee e Toutanova, 2019).

A transformer is a transduction model with an encoder-decoder structure, although it uses self-attention to compute representations of its input and output without using sequence aligned RNNs or convolution (Figure 2.5). "An attention function can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. Self-attention, also called intra-attention, is an attention mechanism relating different positions of a single sequence to compute a representation of the sequence" (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser e Polosukhin, 2017).

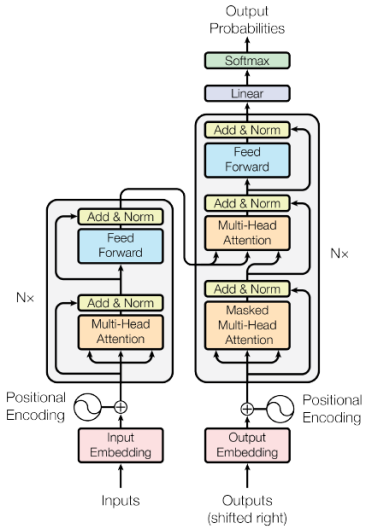


Figure 2.5 - The Transformer model architecture (Devlin, Chang, Lee e Toutanova, 2019)

BERT has a multi-layer bidirectional Transformer encoder architecture and can be found in two sizes BERT Base (12 layers, hidden size = 768, 12 self-attention heads, total parameters=110M), comparable to OpenAI GPT, and BERT Large (24 layers, hidden size = 1024, 16 self-attention heads, total parameters = 340M). During the pre-training phase, WordPieces embeddings with 30,000 vocabulary tokens are used and two unsupervised tasks are applied: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). NSP is applied in order to the model capture the relationship between two sentences. This task is especially important to improve results in tasks like Question Answering and Sentence Pair Classification (Devlin, Chang, Lee e Toutanova, 2019).

The fine-tuning phase is comparatively faster than the pre-training, even though all the network weights are fine-tuned end-to-end. Also, the input and outputs can vary for each downstream task. For instance, in Sentence Classification, only the first token (the special token [CLS]) is passed to the single-layer classifier, while in Sentence Tagging, every single token is passed to the top layer, as represented in Figure 2.6 (Devlin, Chang, Lee e Toutanova, 2019).

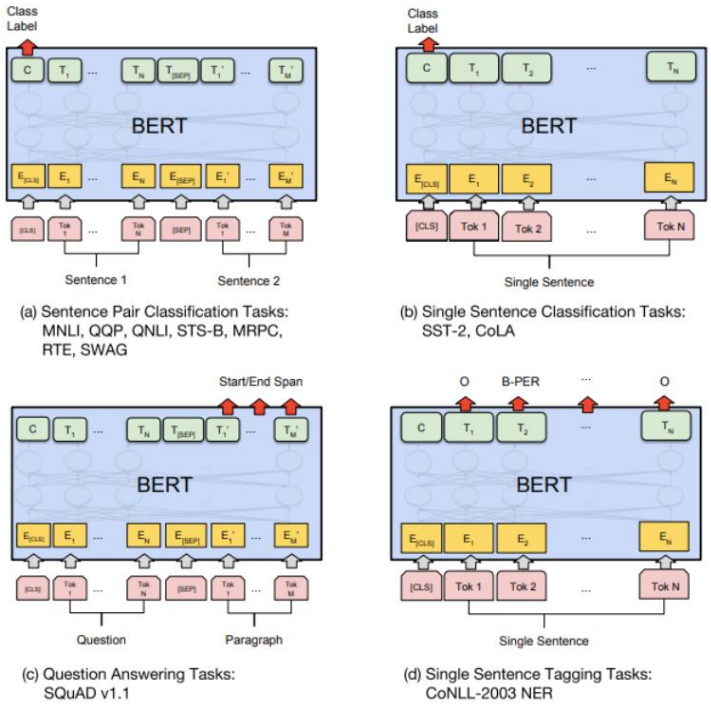


Figure 2.6 - Illustrations of fine-tuning BERT on different tasks (Devlin, Chang, Lee e Toutanova, 2019)

In addition to a large dataset, training a model like this requires substantial infrastructure and takes considerable time. On the other hand, model fine-tuning is a faster and less resource-consuming activity. Therefore, having pre-trained models available becomes advantageous, especially for text mining use cases with a small dataset or without high computational resources available, since it allows them to be solved from a robust LM (Howard e Ruder, 2018).

Initially, English and multilingual BERT pre-trained models from both size and case/uncased versions were made available. Then, many other models were released. Some are based on the original BERT, like RoBERTa (Liu, Ott, Goyal, Du, Joshi, Chen, Levy, Lewis, Zettlemoyer e Stoyanov, 2019), SpanBERT (Joshi, Chen, Liu, Weld, Zettlemoyer e Levy, 2020), ALBERT (Lan, Chen, Goodman, Gimpel, Sharma e Soricut, 2020), DistilBERT (Sanh, Debut, Chaumond e Wolf, 2019). Others were presented with different approaches and purposes, like Tranforme-XL, XLNet, T5, ERNIE, Electra and GPT-3. Hence, all of them to build a general model that achieves the state-of-the-art on NLP downstream tasks and that could be reused to solve text mining problems.

Once the code of those models was released, versions from different languages and even specific domains were pre-trained and made available as well: the French CamemBERT (Martin, Muller, Suárez, Dupont, Romary, la Clergerie, de, Seddah e Sagot, 2020), the Italian ALBERTO (Polignano,

Basile, Gemmis, de, Semeraro e Basile, 2019), the Spanish BETO (Cañete, Chaperon, Fuentes e Pérez, 2020), the Dutch BERTje (Vries, Cranenburgh, Bisazza, Caselli, Noord e Nissim, 2019), bioBERT for biomedical language representation (Lee, Yoon, Kim, Kim, Kim, So e Kang, 2020) or SciBERT for scientific text (Beltagy, Lo e Cohan, 2019), among others. Similarly, BERTimbau was pre-trained using a large Portuguese Corpus - BrWaC (Brazilian Web as a Corpus) (Souza, Nogueira e Lotufo, 2020).

2.3. Related Studies

Gomez and Moens conducted a survey about automatic patent classification systems and the use of traditional machine learning for text classification and clustering. Their analysis was done with a complete explanation of patents and the classification system. Furthermore, they compared studies with several algorithms, distinct feature processing methods and different levels of classification in IPC. They conclude that patent classification is a complex problem which there were still many alternatives to explore (Gomez e Moens, 2014).

After defining "Intellectual Property Analytics (IPA) as the data science of analyzing a large amount of Intellectual Property information, to discover relationships, trends and patterns for decision", Aristodemou and Tietze reviewed 57 articles and promoted a discussion about AI, machine learning and deep learning approaches to analyze IPA data. They observed the growth of publications over the years, with most of them concentrated in computer science subjects. Moreover, they call attention to the high concentration of articles around artificial neural networks (ANN) and the use of backpropagation learning methods, followed by support vector machine (SVM) and conditional random fields (CRF), focused on classification tasks (Aristodemou e Tietze, 2018).

Zhang used SVM to build sub-classifiers for each class of the patents that are combined in a multi-classifier fusion where the final label is selected using an active learning method (Zhang, 2014). Wu et al. proposed a patent classification system based on a self-organizing map (SOM), kernel principal component analysis (KPCA) and SVM, using quality indicators extracted from different parts of the document (Wu, Chang, Tsao e Fan, 2016). Trappey et al. presented a patent document classification method based on neural network and key phrases frequency that claimed to yield average accuracy above 90% in a specific test set. Further, the same authors used an ontology-based artificial neural network to automatically classify and search knowledge documents (Trappey, Hsu, Trappey e Lin, 2006; Trappey, Trappey, Chiang e Huang, 2013).

DeepPatent is a deep learning algorithm proposed by Li et al. that combines word embedding pre-trained on the title and the abstract sections using skip-gram and a Convolutional Neural Network (CNN) with multi-size filters. This approach aimed to overcome the main limitations of the traditional text encoding and machine learning algorithms, like data sparsity, the inability of capturing complex contents and poor performance on large datasets. Their model presented an F1 top 4 score of 55.09% (Li, Hu, Cui e Hu, 2018). A recent study compares the performance of a hierarchical SVM and various neural network models and applies state-of-the-art hyperparameter optimization techniques on a CNN to understand the effects on the model accuracy. With this optimized neural network, they

achieved 55.02% accuracy on the public Wipo-Alpha dataset[6] (Abdelgawad, Kluegl, Genc, Falkner e Hutter, 2020).

Derieux et al. used different approaches to classify the CLEF-IP dataset with English (68%), German (24%) and French (8%) patents. They found different results according to the language. Indeed, classification on German patents was not less than 10 points below English patent classification. Albeit the size of the training set has a significant impact on these results, they impute this difference to the specificities of each language as well (Derieux, Bobeica, Pois e Raysz, 2010).

Concerned about the specificity of the language used in patent applications, Risch and Krestel proposed domain-specific word embeddings. They trained a fastText model on a dataset of more than 5 million patents in English and evaluated it at the WIPO-alpha dataset through a bi-directional Gated Recurrent Units (GRUs). They concluded that domain-specific representations outperform those trained on Wikipedia, but the underrepresented classes still were a challenge in this problem (Risch e Krestel, 2019).

Following the idea of the Australasian Language Technology Association Workshop 2018 that had launched the challenge of fine-tuning a pre-trained language model to classify Australian patents, Lee and Hsiang pre-trained BERT in a dataset of approximately 3 million documents from Google Patents Public Datasets, using the claims section only and compared their results with the DeepPatent ones. They achieve an F1 score of 63.74% in the IPC subclass level (632 labels) (Lee e Hsiang, 2020).

The table below summarizes the studies mentioned above.

Authors	Feature Engineering	Algorithm	Patent application section used	Language
(Trappey, Hsu, Trappey e Lin, 2006)	Key phrases frequency based on TF-IDF	Neural Networks	full document	English
(Derieux, Bobeica, Pois e Raysz, 2010)	Terms extraction and semantic relation	SVM	full document	English, German, French
(Trappey, Trappey, Chiang e Huang, 2013)	Key phrases frequency based on TF-IDF	Ontology-Based Neural Network	full document	English
(Zhang, 2014)	-	SVM	-	English
(Wu, Chang, Tsao e Fan, 2016)	SOM, KPCA	SVM	full document	English
(Li, Hu, Cui e Hu, 2018)	Skip-gram	CNN	title and abstract	English
(Risch e Krestel, 2019)	Domain-specific FastText word embeddings	Bi-directional GRU	title and abstract	English
(Abdelgawad, Kluegl, Genc, Falkner e Hutter, 2020)	GloVe, Word2Vec, FastText	Hierarchical SVM and CNN with BOHB (Bayesian Optimization hyperband)	title, abstract, description, and claims	English

[6] https://www.wipo.int/classifications/ipc/en/ITsupport/Categorization/dataset/index.html

(Lee e Hsiang, 2020)	-	BERT-Base	claims	English

Table 2.1 - Patent classification related studies

Regardless of not having found any study about patent classification in Portuguese, there are several studies about Portuguese text classification using machine learning algorithms (Gonçalves, Silva, Quaresma e Vieira, 2006). In the same way, pre-trained word-embeddings have been used in other text mining use cases (Castro, Silva e Soares, 2018; Rodrigues, Rodrigues, Castro, de, Silva, da e Soares, 2020). To perform a Named entity recognition (NER) task, Santos and Guimarães pre-trained word-level embeddings using word2Vec in a *corpus* composed by Portuguese Wikipedia, CETENFolha[7] and CETEMPúblico[8] documents (Santos e Guimarães, 2015). An ELMo Language Model trained using a *corpus* created from Portuguese Wikipedia and public documents from Brazil's Labor Courts were provided by Castro, Silva and Soares (Castro, Silva e Soares, 2019). Souza, Nogueira and Lotufo claim to present the first study where BERT models were applied to NER task in Portuguese. The models that they made publicly available can be used in many other NLP tasks in Portuguese (Souza, Nogueira e Lotufo, 2020). The cited authors and their contributions are briefly described in the table below.

Authors	Contribution
(Gonçalves, Silva, Quaresma e Vieira, 2006)	Portuguese text classification using part-of-speech and SVM
(Santos e Guimarães, 2015)	Portuguese word embeddings using word2vec
(Castro, Silva e Soares, 2018)	NER using pre-trained Portuguese word-embeddings
(Gonçalo Oliveira, 2018)	Portuguese word embeddings using Node2Vec
(Castro, Silva e Soares, 2019)	Portuguese ELMo Language Model
(Rodrigues, Rodrigues, Castro, de, Silva, da e Soares, 2020)	Semantic similarity using pre-trained Portuguese word embeddings and pre-trained ELMo Language Model
(Souza, Nogueira e Lotufo, 2020)	Portuguese BERT model

Table 2.2 - Studies using Portuguese corpus

[7] https://www.linguateca.pt/cetenfolha/

[8] https://www.linguateca.pt/cetempublico/

3. METHODOLOGY

To achieve the work project goals, we followed typical text classification solutions (Kowsari et al., 2019). Firstly, the dataset was created based on the .xml files with granted patents made available by INPI. Then, some exploratory analysis was done to understand the dataset.

Since the intention was to explore and compare different algorithms to solve the problem, the two subsequent phases, feature engineering and modelling, followed three distinct experiment paths:

1. Distinct methods to vectorize the patents used as input to machine learning models.
2. Deep learning models using an embedding layer and FastText pre-trained word-embedding.
3. A built-in language model is used to represent the words, and pre-trained models are fine-tuned.

In the final phase, the models' performance was evaluated and compared.

Figure 3.1 brings the blueprint of the methodology used in this work study. The phases will be detailed in this chapter.

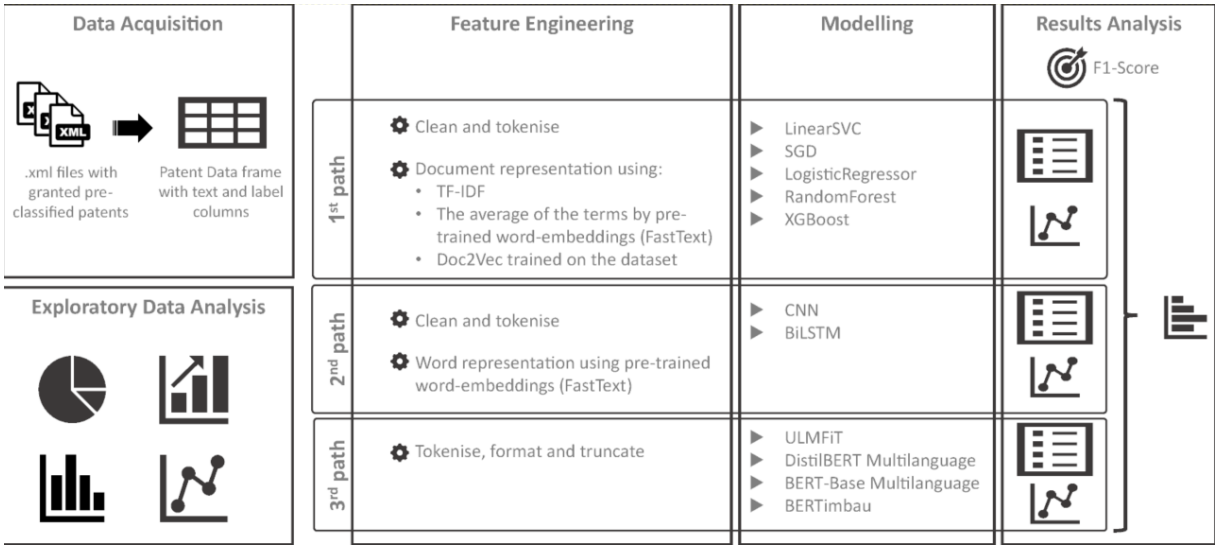


Figure 3.1 – Applied Methodology

3.1. DATA ACQUISITION

A data set of granted pre-classified patents were made available to the participants of the contest promoted by INPI, and the goal was to present a solution to classify patents in the second level of IPC.

There were two types of .xml files with general information of granted patents manually classified. The first one with bibliographic data (patent id, applicants, inventors), title, abstract, IPC codes and CPC codes. The second one with patent id, description and claims. After joining the data, it was

counted approximately 40,000 instances and 124 different categories considering the IPC second level and only the main classification.

The first step was to exclude those instances that have no abstract, claims and description data. Secondly, duplicated patents were identified. It was considered duplicated instances ones with identical id. In these cases, only the first instance was kept to preserve the main IPC classification code. Then, the IPC code was split into 5 columns according to its levels. Finally, a new feature was created by the concatenation of the title and claims or abstract features (Table 3.1).

Feature	Description
id	Patent internal identification
Text	Title (descriptive name of the patent) + Claims (the legal scope of the invention, including delimitations and application field) or Abstract (a brief description of the invention presented in the patent)
Section	IPC 1st level classification code
Class	IPC 2nd level classification code
Subclass	IPC 3rd level classification code
Main group	IPC 4th level classification code
Subgroup	IPC 5th level classification code

Table 3.1 – Dataset features

Many related studies have used title and abstract as input to the classification model. However, more than 80% of the instances in this dataset have no value for abstract while claims and description features have about 0.3% of missing values and all the instances have title information. Therefore, claims were chosen to be used as the input to the classification model. Further, it brings important but more concise information about the patent application than the description. For those instances without claims, abstract data was used. The final dataset had 36,100 instances and no missing values.

3.2. Exploratory Analysis

The dataset contains patents since 1995, but the concentration of data becomes significant from 2005. The most representative years are 2007 and 2013, each with about 11% of the data (Figure 3.2).

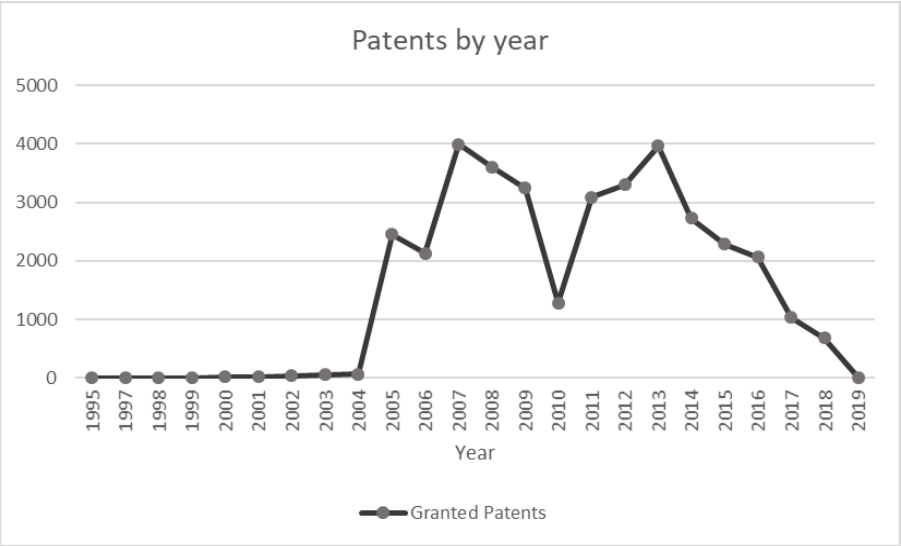


Figure 3.2 - Percentage of patents by year

In Figure 3.3 we can observe that all the eight sections from IPC are present in the dataset but not equally. Sections A and C together count more than 50% of the data, followed by section B (16% approximately). The distribution across the remaining sections is more similar, except for section D with only 646 instances (1,79%).

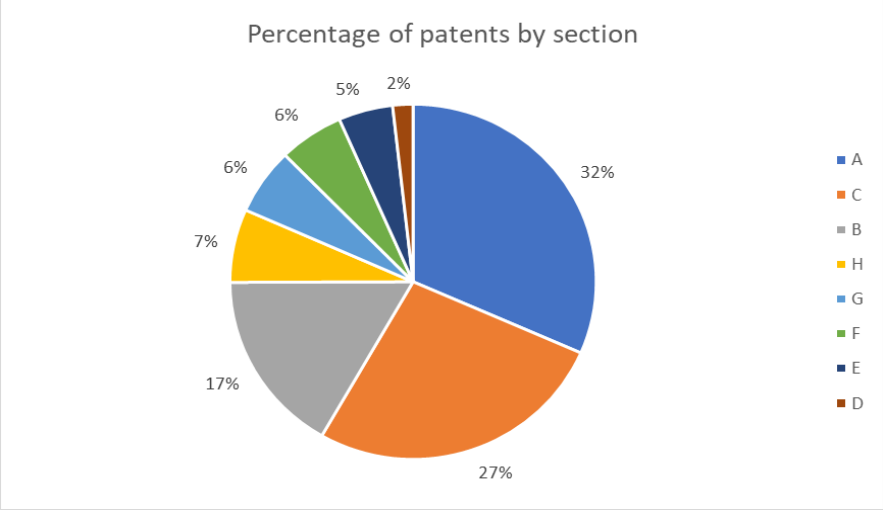


Figure 3.3 - Percentage of patents by section

The unbalance keeps on the class level as well, either by the number of classes on each section or by the number of patents in the classes. In Figure 3.4 we can see the number of classes (IPC 2nd level) per section. Section B has almost two times more classes than C, the second in the rank. Then, A, G and F have a similar number of classes. Regardless of section A has only 16 classes, the most

populated class is A61 that represents 21% of the dataset. Section H has only 5 classes but H04 is the 4th more populated class of the entire dataset, as shown in Figure 3.5.

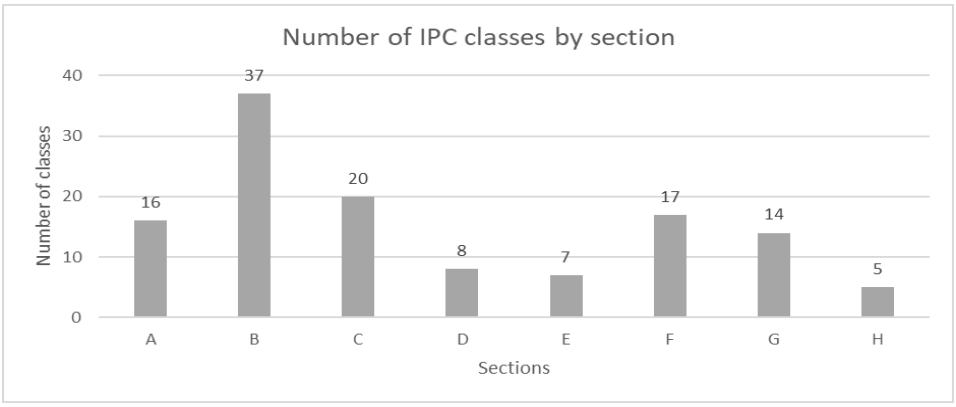


Figure 3.4 - Number of IPC classes by section

Moreover, about 18% of the classes have no more than 20 patents each, 225 training samples (0.6 % of the dataset). Most of the classes have between 20 and 120 instances each. Two classes have more than 2000 patents and together they count 36% of the training dataset, as we can see in Figure 3.6.

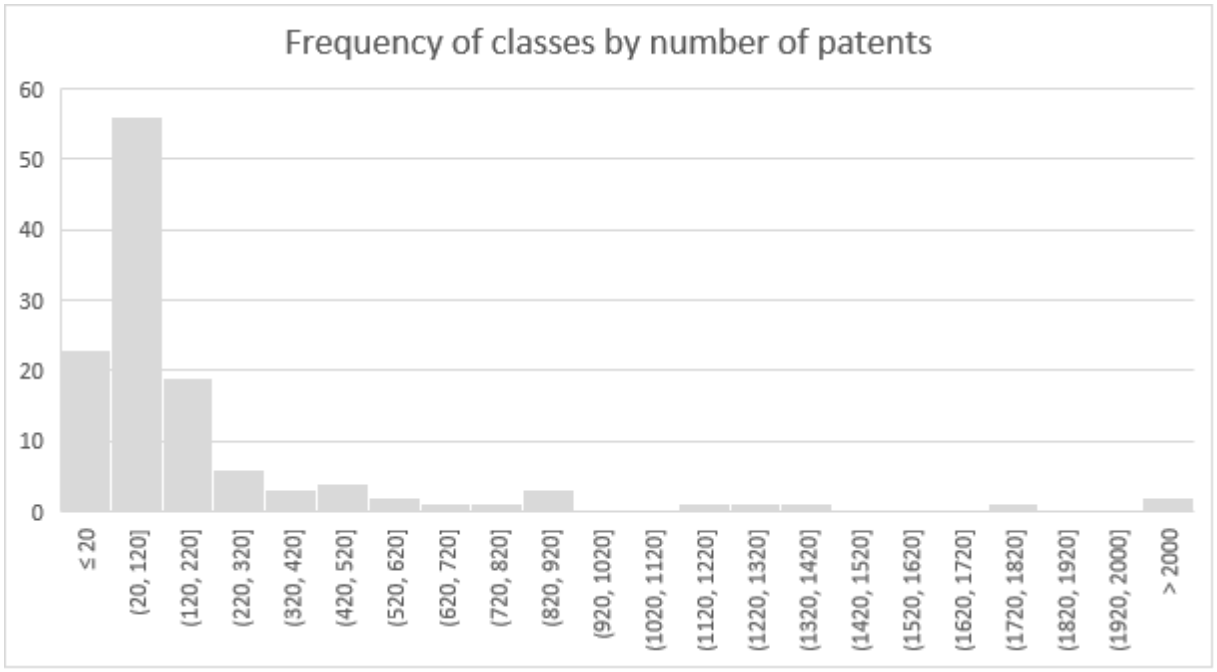


Figure 3.5 – Frequency of classes by the number of patents

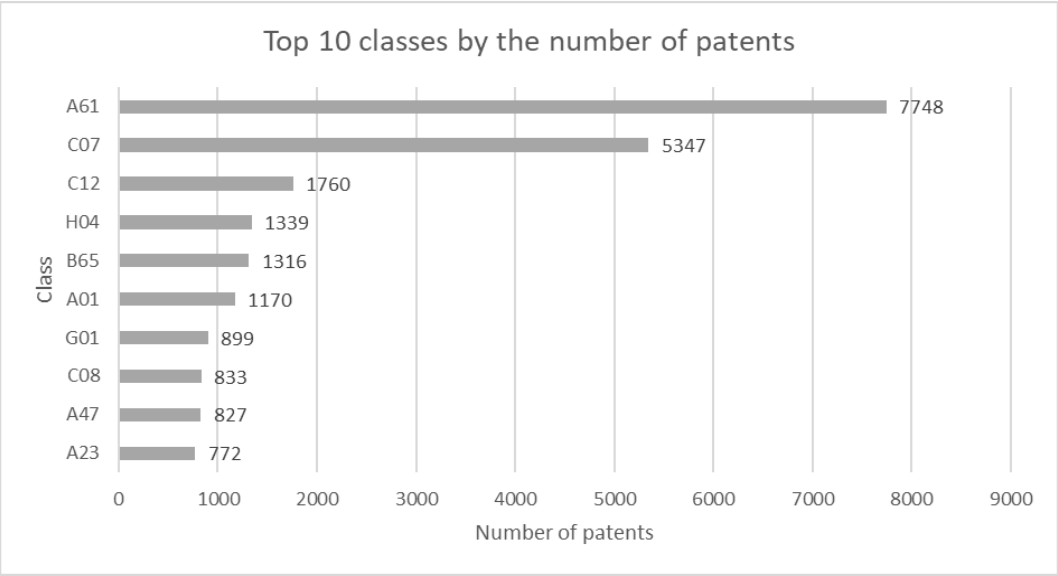


Figure 3.6 - Top 10 classes by the number of patents (IPC 2nd level)

The size of the text that will be used as input to the classification model has some variation as well. After removing stopwords and tokenizing the text, the shorter patent has two tokens, and the longer one has 30,515 tokens. The median text size is 310 tokens, and 96% of the patents have 1,200 tokens at most. Sections C has the biggest text size standard deviation, followed by section H, as shown in the figures below.

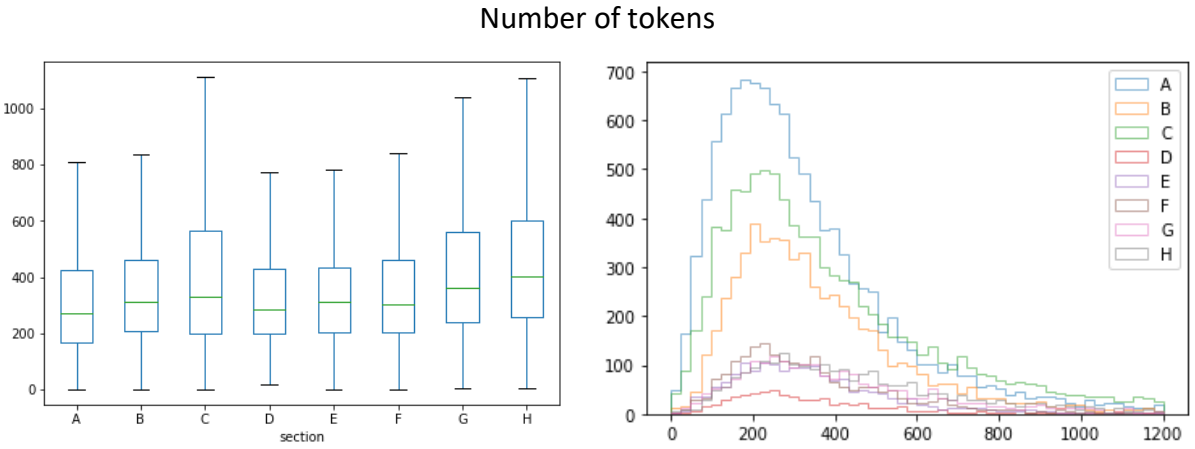


Figure 3.7 – a) Boxplot by section and b) Distribution by section

In Figure 3.8, we can observe the most representative words for each section.

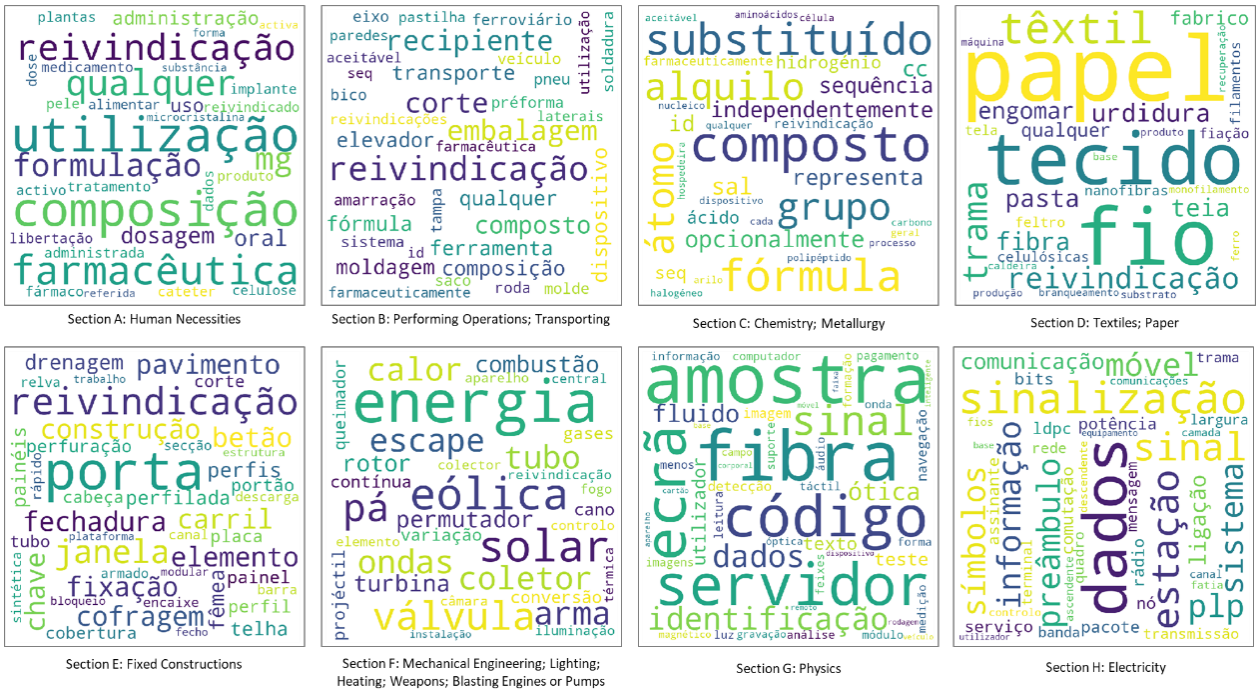


Figure 3.8 – Wordcloud with the more frequent words by section

3.3. Feature Engineering

Feature engineering is a critical step in text mining tasks. The goal is to choose the best way to represent the words of the text to be able to apply the algorithms (Kowsari, Meimandi, Heidarysafa, Mendu, Barnes e Brown, 2019). Different transformations were applied according to the experiment path.

The first approach was to apply TF-IDF to represent the features. To do so, we started cleaning the text and deleting stopwords. Then the cleaned text was tokenized and vectorized.

Even knowing that specific domain-language embeddings bring better performance (Risch e Krestel, 2019), as our dataset is not big enough to train a reasonable Language Model, we decided to validate pre-trained Portuguese word embeddings. We explored word vectors that have been trained in Portuguese corpus, using Word2Vec, Glove, and FastText architectures, on both CBOW and skip-gram strategies. Most of them were well succeeded in recognizing the vocabulary used in the patent dataset. Specifically, FastText models can represent words that are not in the original dataset. For the others, in general, the worst performance occurred in section C (Figure 3.9). This can be explained by the section subject (Chemistry and Metallurgy) and the technical words used to describe these kinds of inventions.

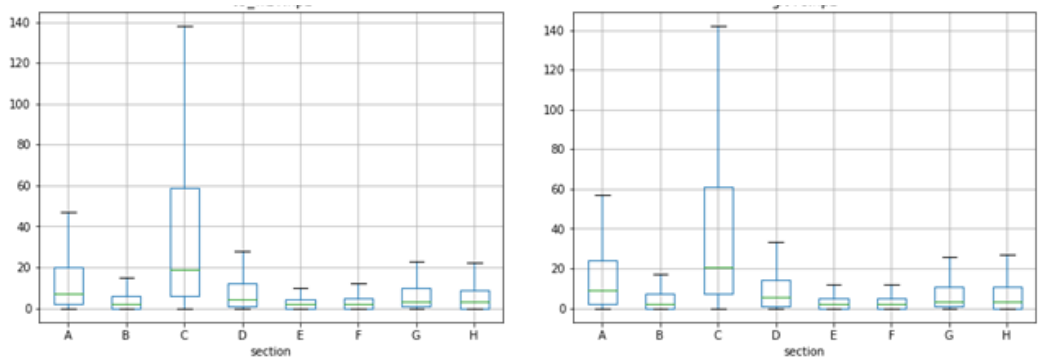


Figure 3.9 - Frequency of out-of-vocabulary words by section using different pre-trained embeddings - Word2Vec and Glove, respectively

So, as a second approach, we again cleaned and tokenized the text and applied FastText[9] word embedding pre-trained on Common Crawl and Wikipedia using CBOW with position-weights, in dimension 300, with character n-grams of length five and window of size 5.

For the pre-trained models, the feature engineering step must be performed using the methods available with the models and the dataset is transformed into the specific format that the model expected. In this case, the data was tokenized and the sentences were prepared with the addition of special tokens. All the documents were truncated to the maximum length of 128 tokens.

3.4. MODELING

Following the paths described previously, traditional machine learning algorithms and ensemble methods were tested, namely Logistic Regression, Linear Support Vector Machine (SVM), Linear model with Stochastic Gradient Descent (SGD), RandomForest, and XGBoost.

Starting with linear classifiers, Logistic Regression is one of the earliest and well-known classification algorithms and the SVM classifier method for many years has outstanding with its effectiveness in text classification. SGD learning allows minibatch and can be a good strategy for large-scale problems. Tree-based classification algorithms, especially voting classifiers like XGBoost, can be fast and accurate for document classification (Kowsari, Meimandi, Heidarysafa, Mendu, Barnes e Brown, 2019).

To start training, after preparing the features by applying TF-IDF, the default algorithms parameters were applied, except by class_weight that was set to "balanced" whenever possible, and the models were evaluated by cross-validation in 5 folds. Following, document embeddings were computed using two different strategies. Firstly, by the mean of the FastText word-embeddings. Secondly, by Doc2Vec, using the whole corpus. The same algorithms were run using the document vectors, and again, a cross-validation strategy was used. The best result was achieved with a LinearSVC model using TF-IDF. Then, a grid search was applied to it as a tuning strategy.

In the second experiment path, two Neural Networks architectures were trained and optimized. A Convolutional Neural Network (CNN) and a bi-directional Long Short-Term Memory (BiLSTM). In both cases, FastText word-embeddings were applied.

[9] https://fasttext.cc/docs/en/crawl-vectors.html

CNN can capture local correlations of spatial or temporal structures. In NLP tasks, it means extracting n-gram features at different positions of text through a series of convolutional filters. These kinds of models have been achieving good performance in text classifications and also for some cases of patent classification (Hu, Li, Hu e Yang, 2018; Li, Hu, Cui e Hu, 2018).

LSTM was designed to handle sequence data and capture long-term dependence while controlling the ratio of information to forget and to store during the training. A BiLSTM is a combination of two LSTM in which the context is seen in both directions, from left to right (forward) and from right to left (backward). It means, for each word, capture previous and following information, with the ability to remember or forget it when necessary. In the end, the weights of the two networks are combined to compute the output. For this property of handle long-term dependencies, BiLSTM has been widely used in text classification (Bispo, Macedo, Santos, Silva, Da, Matos, Prado, Silva, Da e Guimarães, 2019; Devlin, Chang, Lee e Toutanova, 2019; Hu, Li, Hu e Yang, 2018).

In the last experiment path, pre-trained BERT, DistilBERT, and ULMFiT models were applied. We started with BERT-Base Multilingual Cased, a model made available by BERT authors which supports 104 languages. Next, BERTimbau which is a BERT model trained on the BrWaC (Brazilian Web as Corpus)[10], a large Portuguese corpus, for 1,000,000 steps, using the whole-word mask. Then, DistilBERT Base Multilingual Cased, this model is based on BERT architecture, supports the same 104 languages, and uses the BERT tokenizer, but it is built with only 6 layers, half of the BERT-Base model. It aims to be lighter and to run faster, since it has fewer parameters, while preserving a good performance. For BERT models several warmup values, text length, and hyperparameters were tried. Finally, a hyperparameter tuning and unfreeze strategy were applied to ULMFiT.

3.5. ASSESSMENT

To check the performance of the models, the dataset was split into a training set with 25,267 patents and the test set with the remaining 30% of the documents. Once the dataset was high unbalanced, we undersampled the most represented classes (A61 and C07) on the training set but kept all the samples for the test set.

For each experiment path, after training the models, the unseen test set was prepared with the required transformations and submitted to the model to predict the patent class. We used the predicted and real labels to compute Precision, Recall and F1 weighted as evaluation metrics.

Considering True Positive (TP) as the number of positive examples correctly classified, False Positive (FP) the negative examples classified as positive and False Negative (TN) the number of positive examples misclassified, we have:

$$Precision = \frac{TP}{FP + TP} \qquad Recall = \frac{TP}{TP + FN} \qquad F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

F1 weighted means that the score is computed for each label, then the average is calculated and weighted by support (the number of true instances for each label).

[10] https://www.researchgate.net/publication/326303825_The_brWaC_Corpus_A_New_Open_Resource_for_Brazilian_Portuguese

4. RESULTS AND DISCUSSION

In this chapter, the results obtained for the proposed approaches will be presented and discussed. Some discussion about the performance of the best model will be done in the end.

In Table 4.1 we can observe the results for each algorithm and feature engineering method applied in the first path. The best result was with LinearSVC using TF-IDF. Then, a grid search was applied to it as a tuning strategy. The tuned LinearSVC presented an f1-weighted score of 0.608 on the test set and was used as the baseline.

	F1_weighted (mean score)		
Model	**TF-IDF**	**Doc Embedding (Fasttext)**	**Doc Embedding (Doc2Vec)**
LinearSVC	0.586 (+/-0.003)	0.223 (+/-0.005)	0.398 (+/-0.003)
LinearModel + SGD	0.530 (+/-0.005)	0.300 (+/-0.048)	0.374 (+/-0.003)
LogisticRegressor	0.518 (+/-0.004)	0.478 (+/-0.004)	0.447 (+/-0.002)
RandomForest	0.151 (+/-0.012)	0.408 (+/-0.004)	0.044 (+/-0.003)
XGBoost	0.505 (+/-0.006)	0.436 (+/-0.005)	0.337 (+/-0.004)

Table 4.1 - Mean F1 score (cv=5) with different feature engineering methods

We can note that the performance of these models using document embeddings was in general worse than the same algorithms when TF-IDF was used. The strategy of aggregating the terms of the document by the mean did not result in a good tactic. The same word-embeddings were used as input to the Neural Networks with a better result.

The best results for each model in the second and third paths can be seen in Table 4.2. While CNN and DistilBERT got the lowest scores, BERT models achieved the best performance.

Model	**F1_weighted (%)**
LinearSVC (baseline)	60.80
CNN	50.00
DistilBERT Multilingual	50.10
BiLSTM	57.00
ULMFiT	57.00
BERT-Base Multilingual	59.50
BERTimbau	63.60

Table 4.2 - F1 score on the test set

Regardless of they share the grammar, Brazilian Portuguese and European Portuguese present notable differences in vocabulary and sentence structure. Hence it is interesting to note that despite BERTimbau having been trained exclusively on a Brazilian Corpus, it was the model that presented the best performance to solve this problem, even than the multilingual BERT, trained on a corpus

with both variants. However, the performance was only 4% superior to the baseline, despite the complexity of the model.

In the analyses of the results obtained using the pre-trained BERTimbau, when we observe the section level, sections A, C, and H have average F1 score higher than the general F1 score, 17%, 10% e 9% higher, respectively. On the other hand, sections D and G have the worst score by section. (Figure 4.1).

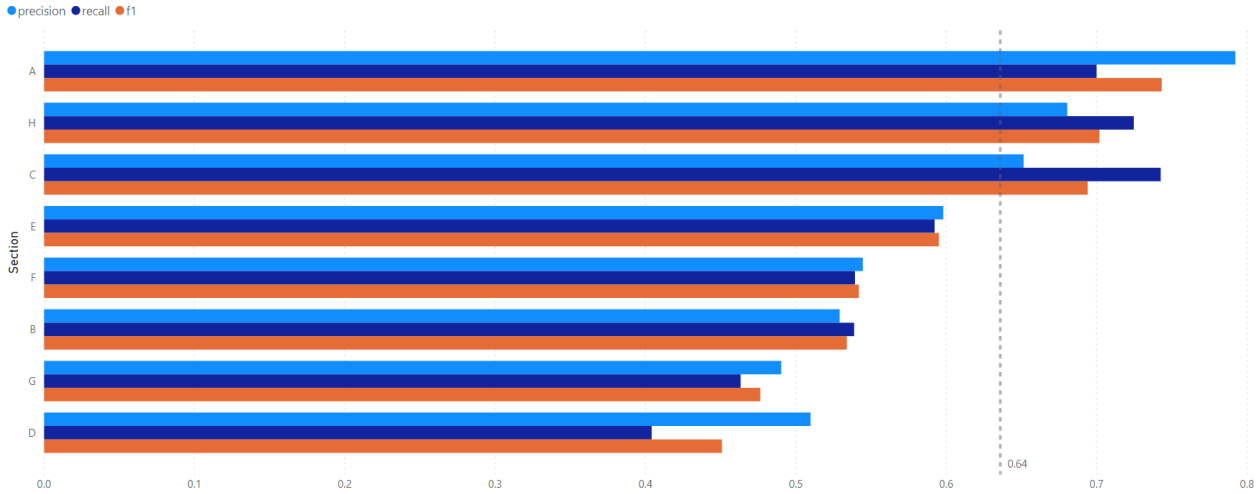


Figure 4.1 - Precision, Recall, and F1 score by section

The worst result in section D was not a surprise since it had the lowest amount of training samples. On the other hand, the more numerous classes - A61 e C07 (2000 samples in training set each) - are in the top 10 best F1 scores, with 79.70% and 80%, respectively. However, the number of training samples was not a factor that decisively affected the results.

In general, the classes with more than 500 training samples had F1 scores bigger than 0.680. But surprisingly, the best result (F1 score = 88.20%) was A24 that had had less than 100 samples on the training set. In the same way, A43 (F1 score = 81.60%) and B64 (F1 score = 77.40%) are presented in the top 10 scores; regardless, they have had only 78 and 41 samples in the training set, respectively.

For those classes, good performance could be explained by their specific context. The more frequent words in class A24, like "*cigarro*", "*tabaco*" and "*filtro*", are highly representative of its context. The same happens in class A43 with "*sola*", "*calçado*", "*palmilha*" or "*aeronave*", "*avião*" and "*aterragem*" for class B64, as shown in Table 4.3 and Figure 4.2

Class	Context
A24	TOBACCO; CIGARS; CIGARETTES; SIMULATED SMOKING DEVICES; SMOKERS' REQUISITES
A43	FOOTWEAR
B64	AIRCRAFT; AVIATION; COSMONAUTICS

Table 4.3 - Classes A24, A43, and B64 classification subject

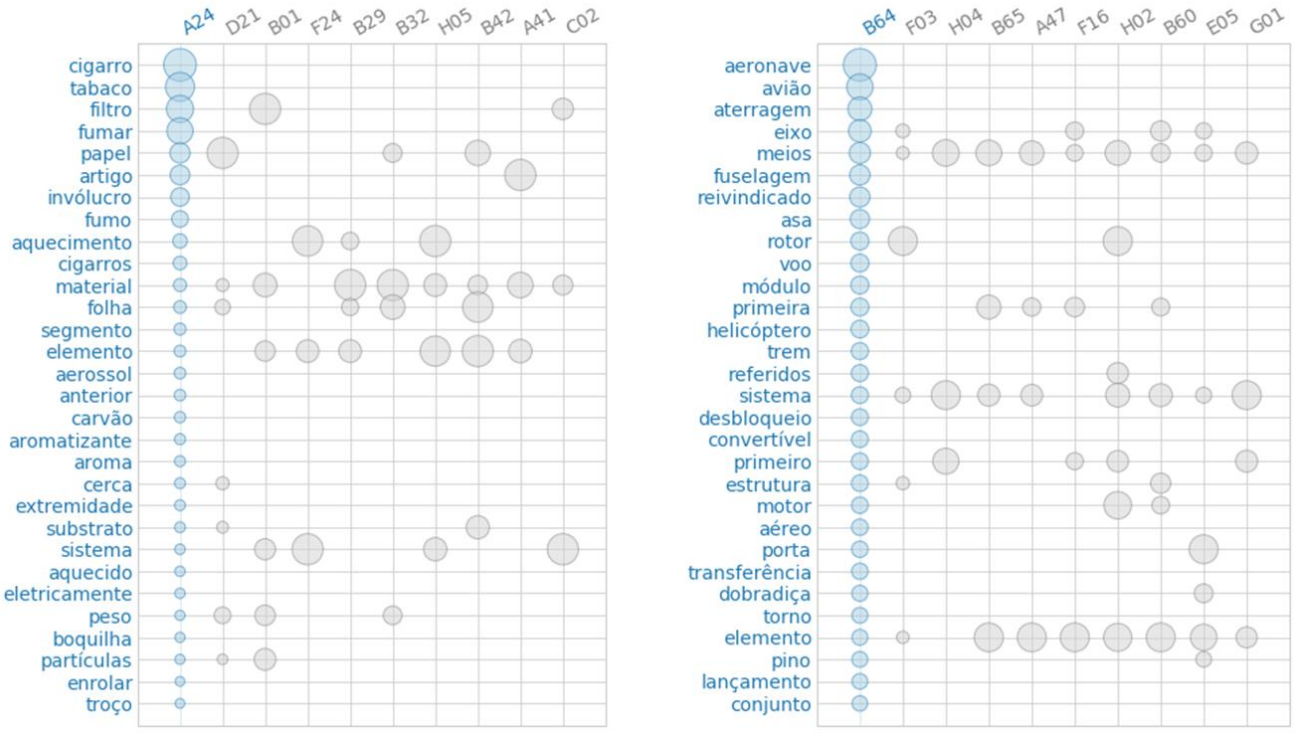


Figure 4.2 - Classes A24 and B64 most frequent words in parallel to most similar classes

In contrast, classes like B01 (431 samples), F16 (407 samples), and G06 (316 samples) scored less than 50% on the test set, even with a considerable quantity of samples in the training phase. In some cases, the broader context of those classes, overlap with other classes or related terms could explain the misclassification.

For instance, 60% of the misclassified instances on class B01 (Physical or Chemical Process or Apparatus in general) were predicted as some class of section C which also deals with the Chemistry subject. Besides, almost 30% of G06 (Computing; Calculating or Counting) instances were classified as H04 (Electric Communication Technique). In **Error! Reference source not found.** we can observe an overlap of terms (*dados*, *sistema*, *informação*, *utilizador*).

Figure 4.3 shows the top frequent words for some of those classes and the similarity of terms when compared to the classes to which they were the most confounded.

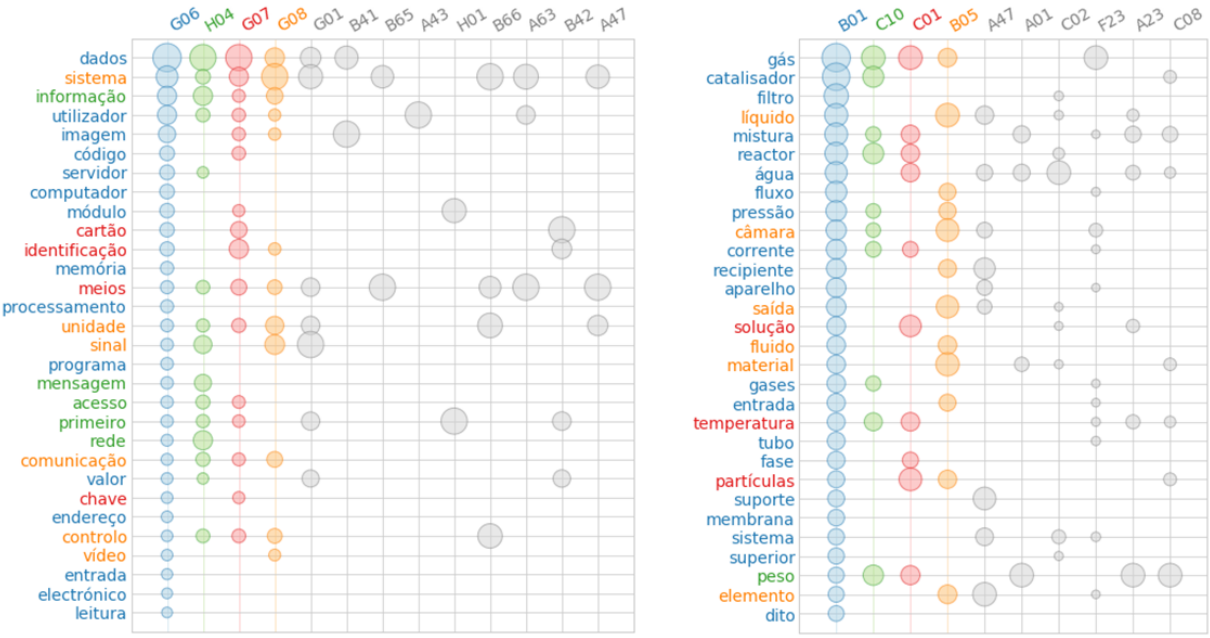


Figure 4.3 - Classes G06 and B01 most frequent words in parallel to most similar classes

Despite the good result in some classes with a small number of samples, in general, BERT did not have a good performance when classes had less than 60 training samples. In this group, 44 classes had an F1 score equal to 0. Although it counts only for 4% of the test set, it represents about 36% of the classes.

The f1 score by class obtained on the test set can be observed in Appendix 8.1.

5. CONCLUSIONS

Patents are more than ever used by companies, not only as a financial protection instrument but also as a database for researches and innovation. Since a patent is a lengthy descriptive document, it also becomes an interesting dataset to be explored in text mining tasks.

This work project aimed to train a model to classify Portuguese patents in the second level of ICP, one of the critical phases of the grant patent process, using a dataset with granted patents put available by INPI. Then, evaluate the results and identify the most relevant attributes in the classification process.

For the feature engineering phase, we explored the TF-IDF method, Fasttext pre-trained word embedding, and we also trained Doc2Vec in the whole dataset. Traditional machine learning algorithms, ensembled methods, and neural networks were used to train the model, with the best performance presented by a LinearSVC model using TF-IDF. Following, some pre-trained models were also explored and tuned on the dataset. After a fine-tuning phase, BERTTimbau - a BERT architecture model pre-trained on a large Portuguese corpus - presented the best results to solve the task with an F1 score of 63.60%. Although, the performance was not much superior to the LinearSVC model used as a baseline.

Since the dataset was high unbalanced, as usual in patent applications, it was expected that the classes with the lowest quantity of samples presented the worst performance. It happened in some cases, especially in classes with less than 60 training samples. However, the number of training samples was not the decisive factor. For instance, the class with the best performance achieved an F1 score superior to 80% with only 117 training samples. The specificity of the context was a relevant factor here. The most frequent words in this class were good representants of its context, like "*cigarro*", "*tabaco*", and "*filtro*". On the other hand, classes on which general words like, "*dados*", "*sistema*" and "*informação*" were broadly used to describe the patent were easily misclassified.

Along with the unbalance, the high number of classes (124) brought a particular challenge to the model that did not classify correctly any patent of 44 classes, almost 36% of the total. Although they contributed only with 4% of the patents in the test set.

Patent classification is a challenging task because of the hierarchical classification system, but also because of the way a patent is described, the overlap of the contexts, and also the underrepresentation of the classes. Even so, we consider the final model presented an acceptable performance given the size of the dataset, the computational resources applied, and the task complexity. In addition, it is an area of growing interest that can be leveraged by the new researches that are revolutionizing text mining.

6. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

A patent application requires the definition of the invention but also its delimitation. The comprehension of the text and its structure is a key point to correctly classify this kind of documents. For this project, we faced limitations related to the size of the dataset and the computational resources, both of them important aspects for training models in text mining tasks. Therefore, a path for future work could be the training of domain-specific word embeddings in a large corpus of Portuguese patents to capture special characteristics of patents description in this language.

Besides, building hierarchical models from the section to the subgroup level could be a way to deal with the challenge of overlapped contexts. Finally, to handle the underrepresentation of some classes, an autoregressive language model like GPT-3, which uses deep learning to produce human-like text, could be explored to oversample those classes.

7. BIBLIOGRAPHY

ABDELGAWAD, Louay; KLUEGL, Peter; GENC, Erdan; FALKNER, Stefan; HUTTER, Frank - Optimizing Neural Networks for Patent Classification. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**. . ISSN 16113349. 11908 LNAI:2020) 688–703. doi: 10.1007/978-3-030-46133-1_41.

ARISTODEMOU, Leonidas; TIETZE, Frank - The state-of-the-art on Intellectual Property Analytics (IPA): A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data. **World Patent Information**. . ISSN 01722190. 55:July (2018) 37–51. doi: 10.1016/j.wpi.2018.07.002.

BELTAGY, Iz; LO, Kyle; COHAN, Arman - SciBERT: Pretrained Contextualized Embeddings for Scientific Text. **arXiv Computer Science**. 2019). doi: arXiv:1903.10676v2.

BISPO, Thiago D.; MACEDO, Hendrik T.; SANTOS, Flávio De O.; SILVA, R. P. DA; MATOS, Leonardo N.; PRADO, Bruno O. P.; SILVA, Gilton J. F. DA; GUIMARÃES, Adolfo - Long short-term memory model for classification of english-PtBR cross-lingual hate speech. **Journal of Computer Science**. . ISSN 15526607. 2019). doi: 10.3844/jcssp.2019.1546.1571.

BOJANOWSKI, Piotr; GRAVE, Edouard; JOULIN, Armand; MIKOLOV, Tomas - Enriching Word Vectors with Subword Information. **Transactions of the Association for Computational Linguistics**. . ISSN 2307-387X. 5:2017) 135–146. doi: 10.1162/tacl_a_00051.

CAÑETE, José; CHAPERON, Gabriel; FUENTES, Rodrigo; PÉREZ, Jorge - Spanish Pre-Trained BERT Model and Evaluation Data. **PML4DC at ICLR 2020**. 2020) 1–10.

CASTRO, Pedro Vitor Quinta De; SILVA, Nádia Félix Felipe Da; SOARES, Anderson Da Silva - Portuguese Named Entity Recognition Using LSTM-CRF. Em **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**

CASTRO, Pedro Vitor Quinta De; SILVA, Nádia Félix Felipe Da; SOARES, Anderson Da Silva - Contextual representations and semi-supervised named entity recognition for Portuguese language. Em **CEUR Workshop Proceedings**

Código de Propriedade Intelectual - [online]. Lisboa : Instituto Nacional da Propriedade Industrial, 2019 Available at <www.inpi.pt>.

Cooperative Patent Classification - [online] [Accessed 12 jul. 2020]. Available at <https://www.cooperativepatentclassification.org/index>.

DERIEUX, Franck; BOBEICA, Mihaela; POIS, Delphine; RAYSZ, Jean Pierre - Combining semantics and statistics for patent classification. Em **CEUR Workshop Proceedings**

DEVLIN, Jacob; CHANG, Ming Wei; LEE, Kenton; TOUTANOVA, Kristina - BERT: Pre-training of deep bidirectional transformers for language understanding. **NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference**. 1:Mlm (2019) 4171–4186.

EPO - Patent Index 2019 - [online] Available at <epo.org/patent-index2019>.

Espacenet - Home page - [online] [Accessed 12 jul. 2020]. Available at <https://lp.espacenet.com/>.

FELDMAN, Ronen; SANGER, James - **The Text Mining Handbook** [online] Available at <www.cambridge.org/core/product/0634B1DF14259CB43FCCF28972AE4382>. ISBN

9780511546914.

GOMEZ, Juan Carlos; MOENS, Marie Francine - A survey of automated hierarchical classification of patents. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**. . ISSN 16113349. 2014). doi: 10.1007/978-3-319-12511-4_11.

GONÇALO OLIVEIRA, Hugo - Learning Word Embeddings from Portuguese Lexical-Semantic Knowledge Bases. Em **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**. ISBN 9783319997216

GONÇALVES, Teresa; SILVA, Cassiana; QUARESMA, Paulo; VIEIRA, Renata - Analysing part-of-speech for Portuguese text classification. Em **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**. ISBN 3540322051

Google AI Blog: Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing - [online] [Accessed 7 jul. 2020]. Available at <https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html>.

HOWARD, Jeremy; RUDER, Sebastian - Universal language model fine-tuning for text classification. Em **ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)**. ISBN 9781948087322

HU, Jie; LI, Shaobo; HU, Jianjun; YANG, Guanci - A hierarchical feature extraction model for multi-label mechanical patent classification. **Sustainability (Switzerland)**. . ISSN 20711050. 10:1 (2018). doi: 10.3390/su10010219.

IP5 Statistics Report - [online] [Accessed 12 jul. 2020]. Available at <https://www.fiveipoffices.org/statistics/statisticsreports/2019edition>.

JOSHI, Mandar; CHEN, Danqi; LIU, Yinhan; WELD, Daniel S.; ZETTLEMOYER, Luke; LEVY, Omer - SpanBERT: Improving Pre-training by Representing and Predicting Spans. **Transactions of the Association for Computational Linguistics**. . ISSN 2307-387X. 2020). doi: 10.1162/tacl_a_00300.

JURAFSKY, Daniel; MARTIN, James H. - Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (second edition). **Speech and Language Processing An Introduction to Natural Language Processing Computational Linguistics and Speech Recognition**. . ISSN 08912017. 21:2008) 1–1044. doi: 10.1162/089120100750105975.

KHURANA, Diksha; KOLI, Aditya; KHATTER, Kiran; SINGH, Sukhdev - Natural Language Processing: State of The Art, Current Trends and Challenges. 2017).

KORDE, Vandana - Text Classification and Classifiers:A Survey. **International Journal of Artificial Intelligence & Applications**. . ISSN 09762191. 3:2 (2012) 85–99. doi: 10.5121/ijaia.2012.3208.

KOWSARI, Kamran; MEIMANDI, Kiana Jafari; HEIDARYSAFA, Mojtaba; MENDU, Sanjana; BARNES, Laura; BROWN, Donald - Text classification algorithms: A survey. **Information (Switzerland)**. . ISSN 20782489. 10:4 (2019). doi: 10.3390/info10040150.

LAN, Zhenzhong; CHEN, Mingda; GOODMAN, Sebastian; GIMPEL, Kevin; SHARMA, Piyush; SORICUT, Radu - Albert: A LITE BERT FOR SELF-SUPERVISED LEARNING OF LANGUAGE REPRESENTATIONS. 2020) 1–17.

LEE, Jieh Sheng; HSIANG, Jieh - Patent classification by fine-tuning BERT language model. **World**

Patent Information. . ISSN 01722190. 61:2020) 1–6. doi: 10.1016/j.wpi.2020.101965.

LEE, Jinhyuk; YOON, Wonjin; KIM, Sungdong; KIM, Donghyeon; KIM, Sunkyu; SO, Chan Ho; KANG, Jaewoo - BioBERT: A pre-trained biomedical language representation model for biomedical text mining. **Bioinformatics**. . ISSN 14602059. 2020). doi: 10.1093/bioinformatics/btz682.

LI, Shaobo; HU, Jie; CUI, Yuxin; HU, Jianjun - DeepPatent: patent classification with convolutional neural networks and word embedding. **Scientometrics**. . ISSN 15882861. 117:2 (2018) 721–744. doi: 10.1007/s11192-018-2905-5.

LIDDY, Elizabeth D. - Natural Language Processing. Em **Encyclopedia of Library and Information Science**. 2nd Ed. ed. NY : Marcel Dekker Inc, 2001

LIU, Yinhan; OTT, Myle; GOYAL, Naman; DU, Jingfei; JOSHI, Mandar; CHEN, Danqi; LEVY, Omer; LEWIS, Mike; ZETTLEMOYER, Luke; STOYANOV, Veselin - RoBERTa: A Robustly Optimized BERT Pretraining Approach. 1 (2019).

MANNING, Christopher D. - Electra : Pre - Training Text Encoders As Discriminators Rather Than Generators. **Iclr**. 2020) 1–18.

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHUTZE, Hinrich - **Introduction to Information Retrieval**. [S.l.] : Cambridge University Press, 2008. ISBN 978-0521865715.

MARTIN, Louis; MULLER, Benjamin; SUÁREZ, Pedro Javier Ortiz; DUPONT, Yoann; ROMARY, Laurent; LA CLERGERIE, Éric Villemonte DE; SEDDAH, Djamé; SAGOT, Benoît - CamemBERT: a Tasty French Language Model. Em **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics** [online]. [S.l.] : Association for Computational Linguistics (ACL), 2020 Available at <https://www.aclweb.org/anthology/2020.acl-main.645.pdf>.

MERITY, Stephen; KESKAR, Nitish Shirish; SOCHER, Richard - An Analysis of Neural Language Modeling at Multiple Scales. 2018).

MERITY, Stephen; KESKAR, Nitish Shirish; SOCHER, Richard - Regularizing and optimizing LSTM language models. Em **6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings**

MIKOLOV, Tomas; CHEN, Kai; CORRADO, Greg; DEAN, Jeffrey - Efficient estimation of word representations in vector space. **1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings**. 2013) 1–12.

MIROŃCZUK, Marcin Michał; PROTASIEWICZ, Jarosław - A recent overview of the state-of-the-art elements of text classification. **Expert Systems with Applications**. . ISSN 09574174. 106:2018). doi: 10.1016/j.eswa.2018.03.058.

NADKARNI, Prakash M.; OHNO-MACHADO, Lucila; CHAPMAN, Wendy W. - Natural language processing : an introduction. 2011). doi: 10.1136/amiajnl-2011-000464.

PAN, Sinno Jialin; YANG, Qiang - A survey on transfer learning. **IEEE Transactions on Knowledge and Data Engineering**. . ISSN 10414347. 22:10 (2010). doi: 10.1109/TKDE.2009.191.

PENNINGTON, Jeffrey; SOCHER, Richard; MANNING, Christopher D. - GloVe: Global vectors for word representation. Em **EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference**. [S.l.] : Association for Computational Linguistics (ACL), 2014. ISBN 9781937284961

PETERS, Matthew E.; NEUMANN, Mark; IYYER, Mohit; GARDNER, Matt; CLARK, Christopher; LEE, Kenton; ZETTLEMOYER, Luke - Deep contextualized word representations. Em **NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference**. ISBN 9781948087278

POLIGNANO, Marco; BASILE, Pierpaolo; GEMMIS, Marco DE; SEMERARO, Giovanni; BASILE, Valerio - AlBERTo: Italian BERT language understanding model for NLP challenging tasks based on tweets. **CEUR Workshop Proceedings**. . ISSN 16130073. 2481:2019).

RISCH, Julian; KRESTEL, Ralf - Domain-specific word embeddings for patent classification. **Data Technologies and Applications**. . ISSN 25149288. 53:1 (2019) 108–122. doi: 10.1108/DTA-01-2019-0002.

RODRIGUES, Ruan Chaves; RODRIGUES, Jéssica; CASTRO, Pedro Vitor Quinta DE; SILVA, Nádia Felix Felipe DA; SOARES, Anderson - Portuguese language models and word embeddings: Evaluating on semantic similarity tasks. Em **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**. ISBN 9783030415044

ROSENFELD, Ronald - Two decades of statistical language modeling: Where do we go form here? **Proceedings of the IEEE**. . ISSN 00189219. 2000). doi: 10.1109/5.880083.

SANH, Victor; DEBUT, Lysandre; CHAUMOND, Julien; WOLF, Thomas - DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. 2019).

SANTOS, Cicero Dos; GUIMARÃES, Victor - Boosting Named Entity Recognition with Neural Character Embeddings. Em **Proceedings of the Fifth Named Entity Workshop** [online]. [S.l.] : Association for Computational Linguistics (ACL), 2015 Available at <https://aclanthology.org/W15-3904>.

SILVA, Catarina; RIBEIRO, Bernardete - **Inductive Inference for Large Scale Text Classification**. Berlin : Springer, 2010. ISBN 978-3-642-04533-2.

SOUZA, Fábio; NOGUEIRA, Rodrigo; LOTUFO, Roberto - **Portuguese Named Entity Recognition using BERT-CRF** [online] Available at <https://arxiv.org/abs/1909.10649>.

TRAPPEY, Amy J. C.; HSU, Fu Chiang; TRAPPEY, Charles V.; LIN, Chia I. - Development of a patent document classification and search platform using a back-propagation network. **Expert Systems with Applications**. . ISSN 09574174. 31:4 (2006) 755–765. doi: 10.1016/j.eswa.2006.01.013.

TRAPPEY, Amy J. C.; TRAPPEY, Charles V.; CHIANG, Tzu An; HUANG, Yi Hsuan - Ontology-based neural network for patent knowledge management in design collaboration. **International Journal of Production Research**. . ISSN 00207543. 2013). doi: 10.1080/00207543.2012.701775.

TRAPPEY, Amy J. C.; TRAPPEY, Charles V.; WU, Chun Yi; LIN, Chi Wei - A patent quality analysis for innovative technology and product development. **Advanced Engineering Informatics**. . ISSN 14740346. 2012). doi: 10.1016/j.aei.2011.06.005.

VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki; USZKOREIT, Jakob; JONES, Llion; GOMEZ, Aidan N.; KAISER, Łukasz; POLOSUKHIN, Illia - Attention is all you need. **Advances in Neural Information Processing Systems**. . ISSN 10495258. 2017-Decem:Nips (2017) 5999–6009.

VRIES, Wietse De; CRANENBURGH, Andreas Van; BISAZZA, Arianna; CASELLI, Tommaso; NOORD, Gertjan Van; NISSIM, Malvina - BERTje: A Dutch BERT Model. 2019).

WIPO - **WIPO Intellectual Property Handbook** [online] [Accessed 3 abr. 2021]. Available at <https://www.wipo.int/edocs/pubdocs/en/wipo_pub_450_2020.pdf>.

WU, Jheng Long; CHANG, Pei Chann; TSAO, Cheng Chin; FAN, Chin Yuan - A patent quality analysis and classification system using self-organizing maps with support vector machine. **Applied Soft Computing Journal**. . ISSN 15684946. 2016). doi: 10.1016/j.asoc.2016.01.020.

ZHANG, Xiaoyu - Interactive patent classification based on multi-classifier fusion and active learning. **Neurocomputing**. . ISSN 09252312. 2014). doi: 10.1016/j.neucom.2013.08.013.

ZHUANG, Fuzhen; QI, Zhiyuan; DUAN, Keyu; XI, Dongbo; ZHU, Yongchun; ZHU, Hengshu; XIONG, Hui; HE, Qing - A Comprehensive Survey on Transfer Learning. **Proceedings of the IEEE**. . ISSN 15582256. 109:1 (2021). doi: 10.1109/JPROC.2020.3004555.

8. APPENDIX

8.1. APPENDIX 1 - F1 SCORE BY IPC SECTION AND IPC CLASS