République Algérienne Démocratique et Populaire
Ministère de l’enseignement supérieur et de la recherche scientifique
Université 8Mai 1945 – Guelma
Faculté des sciences et de la Technologie
Département d’Electronique et Télécommunications

Mémoire de fin d'étude
pour l’obtention du diplôme de Master Académique

Domaine : **Sciences et Technologie**
Filière : **Télécommunications**
Spécialité : **Réseaux et Télécommunications**

Automatic Algerian offensive language detection in social media networks

Présenté par :
BOUCHERIT Oussama

Sous la direction de :
Dr. KHEIREDDINE Abainia

Juillet 2021

Abstract

This research focuses on the detection of offensive and abusive content in Facebook comments in Algerian dialect Arabic, we created a corpus from scratch with over 8.7k texts written in both Arabic and Roman scripts, and annotated with three categories offensive, abusive or normal. We performed a series of automatic classification tests with state-of-the-art algorithms. In addition to rule-based classification with an identification algorithm. The results showed an acceptable performance, but the identification algorithm can still be improved with further investigation.

Chapter 1: Computational Linguistics

Chapter 2: Offensive Language

Chapter 3: Methodology

Chapter 4: Results and experiments

List of Figures

List of Tables

List of acronyms

AI	Artificial Intelligence
ALPAC	Automatic Language Processing Advisory Committee
ANN	Artificial Neural Networks
BCE	Before the Common Era
BiLSTM	Bidirectional Long Short-Term Memory
BOG	Bag Of Words
CA	Classical Arabic
CE	Common Era
CL	Computational Linguistics
CNN	Convolutional Neural Network
DL	Deep Learning
GRU	Gated Recurrent Unit
HTML	Hyper Text Markup Language
ID	IDentification
IR	Information Retrieval
LSTM	Long Short-Term Memory
MIT	Massachusetts Institute of Technology
ML	Machine Learning
MLP	Multi Layer Perception
MSA	Modern Standard Arabic
MT	Machine Translation
NB	Naïve Bayes
NLP	Natural Language Processing
OGTD	Offensive Greek Tweet Dataset
POS	Part Of Speech
SGD	Stochastic Gradient Descent
SVC	Support Vector Classifier
SVM	Support Vector Machines
TF-IDF	Term Frequency Inverse Document Frequency

General Introduction

General Introduction

Nowadays, social media became an important part of people's lives, where they made communications much easier. People across the world can talk and interact using social media platforms, and they can also express their opinion and share other people’s opinion. But this ease of communication comes with serious risks, like the use of hate language, abusive content, cyberbullying, and the use of strong and aggressive language with the intention of offending or threatening a person or a group of people. This kind of behavior endangers people's mental and physical wellbeing and may even lead to depression, and because people spend so much time on social media platforms, and they are more vulnerable to these risks.

To prevent fromthese risks, several researches have been conducted to automatically detect offensive language. There are several approaches in Latin languages (especially in English), and recently some approaches were carried out in official Arabic language, but there is a scarcity of researches in Dialectal Arabic (especially in Algerian dialect). The latter fact may be due to the lack of resources, complexity and ambiguity of the language, which makes processing Algerian dialect a difficult task.

Facebook is the social media platform that is widely used in Algeria, with Algerian Dialect as the mainly used language, which makes it a good source for collecting an Algerian Dialect database. Posts in topics like politics and football, and posts with targeted humor or controversial opinions, contain a lot of comments with offensive and abusive language. Our work consists of detecting offensive language in Facebook comments, with the goal of preventing such behavior and protecting people's wellbeing.

In this work, we focus on classifying comments to three categories: offensive, abusive and normal, where these comments were collected from posts in Facebook public groups and pages. The total crawled data was 10,258 comments, and after the annotation of filtering the comments the size was reduced to 8.7k comments. We have used machine learning and deep learning classifiers for the classification task, as well as a rule-based algorithm that we have created specifically for the Algerian Dialect. For the latter, we have conceived two lexicons, i.e. one for offensive language and another for abusive language. For the state-of-the-art classifiers, we have used the SVM, Gaussian NB, Multinomial NB, CNN, BiLSTM and FastText. The experimental results showed that machine learning classifier maintain high accuracies in contrast to others, as well as our proposal is promising and could be improved further to increase the accuracy. However, in overall, the best reached accuracy (0.77) was acceptable, but was not enough comparing to other text categorization fields.

Chapter I :

Computational

Linguistics

1. Introduction

Computational linguistics (CL) is an interdisciplinary field that regroups linguistics, computer science and logic [1], where its main goal consists of making computer systems able to understand and generate natural language.

This chapter starts with an overview about artificial intelligence, machine learning and computational linguistics. Subsequently, we highlight a brief history and present the major challenges of CL. In addition, we discuss the CL fundamentals and its applications as well. At the end, we state some common approaches of text categorization.

2. Background

2.1. Artificial Intelligence

Artificial Intelligence (AI) is one of the computer sciences fields, where its definition changes over time because of the rapid development and the remarkable progress. Among the recent definitions of AI, we find that it consists of imitating the human intelligence behavior by creating machines that simulate the human behaviors, thinking and decisions [2]. In particular, it is focused on studying how the human brain works, resonates, plans, makes decisions, represents knowledge, learns through the environment.

AI systems can be classified into four categories [2] such as thinking like humans, acting like humans, thinking rationally and acting rationally. It has a various application in different fields to help to solve complex problems with an efficient way, where it could be used in healthcare, education, social media, robotics, finance, etc.

2.2. Machine Learning

One of the most important fields of AI is machine learning (ML), where the machines are designed to be able to learn and be trained with an amount of data. Once a machine is trained with an ensemble of data and created a model, it can make predictions on unkown data instances (i.e. never seen before). More specifically, ML involves feeding data to a computer, and the latter will learn how to grow better using statistical techniques and mathematical models [3]. On the other hand, ML relies on the data availability and does not need any rule-based programming.

Among the ML applications, data mining is one of the most important applications. In overall, it employs the same set of features to represent every occurrence in any dataset [3]. The features may be continuous, categorical, or binary. There are two categories of ML algorithms, i.e. unsupervised and supervised. If data instances are fed to the system with known labels (known outputs), the process is called supervised learning. However, the unsupervised learning, on the other hand, consists of feeding uncategorized data (i.e. unlabelled data) to regroup them into clusters or classes (generally called *clustering*).

Reinforcement learning is another kind of machine learning. The environment (external trainer) provides the training information to the learning system in the form of a scalar reinforcement signal, which is a measure of how well the system works [3]. The learner is not informed which actions to do, but it must figure out which activity results in the best reward by attempting each action independently (one at a time).

2.3. Computational linguistics

The construction of cognitive computers with which humans may freely converse in their native language is the core objective of a future-oriented computational linguistics [4]. This task will, in the long term, assure the creation of a functional theory of language, an objective technique of verification, and a diverse set of practical applications [5].

2.3.1. Natural language processing

Natural language processing (NLP) is an artificial intelligence area that aids computers in understanding, interpreting, and manipulating human language. In order to bridge the gap between the human communication and the machine understanding, NLP draws on a variety of fields, including computer science and linguistics [6]. NLP uses statistical, machine learning, and deep learning models with a computational rule-based modeling of human language. These technologies work together to enable computers to interpret and process human language in the form of text or speech data [7].

2.3.2. Information retrieval

Information retrieval (IR) is a field that aims to create an efficient computer-assisted access to any sort of textual information. IR can be defined as a software program that organizes, stores, retrieves, and evaluates data from document repositories (particularly textual data) [8]. IR system helps users to obtain the information they need, but it does not provide explicit answers to the question.

Indexing and matching are the two main operations in retrieval systems. The first one involves the selection of relevant terms representing the text, while the second (i.e. matching) consists of measuring the similarity between two text representations [9].

3. History

CL is a substantial field of intellectual endeavor as one could find. It was started in 1949, when *Warren Weaver* produced his famous memorandum speculating on the possibility of machine translation.

The first machine translation conference was held at MIT in 1952, and the first journal, Mechanical Translation, was published in 1954. However, it was not until the mid of 60s that the term "*computational linguistics*" became popular. Mechanical Translation and Computational Linguistics was the journal's new name in 1965 [1]. The adoption of the magazine by the Association for Machine Translation (MT) and Computational Linguistics, which was founded in 1962, corresponded with this transition.

David Hays, a member of the National Academy of Sciences' Automatic Language Processing Advisory Committee, is thought to have invented the term "computational linguistics" [1]. The publication of the ALPAC report, which proposed that machine translation be abandoned as a short-term technical aim in favor of more fundamental scientific research in language and language processing, was undoubtedly one of the most dramatic occasions in the field's history. Hays anticipated this and concluded that if money from machine translation could be directed to a new field of study, the most important demand was that the area be given a name [1].

The name stuck, but redirection of the funds did not. MT and CL was succeeded by the American Journal of Computational Linguistics in 1974, which was first only available in microfilm form. This evolved into Computational Linguistics around 1980, and it is still going strong today [1]. Machine translation became more realistic again in the 80s, at least for some people and for some purposes, and the circle was closed in 1986 with the release of the inaugural issue of Computers and Translation, which was renamed Machine Translation in 1988 [1]. the International Journal of Machine Translation followed in 1991. *Warren Weaver*'s notion of machine translation sprang from his time as a cryptographer during World War II, and he saw the problem as one of handling textual content using statistical approaches. However, the founders of Computational Linguistics were mostly linguists, not statisticians, and they saw the computer's potential in carrying out the minutely specified rules that they would write, rather than in deriving a characterization of the translation relation from emergent properties of parallel corpora [1].

Syntactic Structures (1957) by *Noam Chomsky* strengthen the idea of grammar as a logical system, which looked ideally suited to computer applications. The fact that *Chomsky* himself saw little value in such an endeavor, or that the precise framework of axioms and rules he recommended was unsuitable for automatic text analysis, did nothing to lessen the appeal of the general concept

[1]. As a result, CL evolved into a project aimed at developing and implementing formal systems that are increasingly considered as the foundation of linguistic theory. If there is a single event that defines the field, it is undoubtedly *John Cocke*'s proposal in 1960 of the Cocke-Kasami-Younger algorithm for deriving all analyses of a string from a grammar of binary context-free rules [1].

More powerful formalisms were quickly needed to suit the special needs of human language, therefore more generic chart parsers, augmented transition networks, unification grammars, and a variety of other formal and computational devices were developed. As a result, CL evolved into a project aimed at developing and implementing the formal systems that are increasingly considered as the foundation of linguistic theory.

4. Computational linguistics challenges

Several limitations and problems face the computational linguistics field as follows:

A. Contextual words, phrases and homonyms

The same words and phrases can have distinct meanings depending on the context of a statement. Otherwise, several terms sound the same, but have completely distinct meanings.

B. Synonyms

Synonyms can present challenges comparably to contextual understanding. Furthermore, some of these terms may have identical meanings, but others may have varying degrees of complexity, and different persons employ synonyms to signify somewhat different meanings within their vocabulary.

C. Irony and sarcasm

Irony and sarcasm are problematic for machine learning models, because they frequently include words and phrases that, strictly by definition, might be positive or negative, but actually convey the opposite [10].

D. Ambiguity

Ambiguity in CL refers to sentences and phrases that potentially have two or more possible interpretations. The ambiguity could be lexical, semantical, or syntactical.

E. Errors in text and speech

Text analysis might be hampered by misspelled or misinterpreted words. Although autocorrect and grammar checkers can handle frequent errors, unfortunately they don't always comprehend what the writer is trying to say. Mispronunciations, accents, stutters, and other factors might make it challenging for a machine to understand spoken language. These concerns can be mitigated when language databases expand and smart assistants are educated by specific users [11].

F. Colloquialisms and slang

Informal words, expressions, idioms, and cultural slang provide a variety of challenges for CL, particularly for models designed for widespread use. More specifically, colloquialisms (i.e. dialects) are not formal language, they do not have standard orthography, where geographic places may have distinct meanings for the same word. For instance, the word 'كعبة' 'ka3ba' in the capital of Algeria and its suburbs means 'Foolish' but in eastern regions of Algeria it means 'a single piece'.

Furthermore, because cultural slang is always evolving and increasing, new words appear every day. This is why custom model training and regular updates can be beneficial, however it often requires a large amount of data.

5. Computational linguistics fundamentals

5.1. Phonology

Phonology is a linguistics branch dealing with the systematic organization of sound. Phonology is derived from the Ancient Greek prefix "*phono*", which refers to voice or sound, and the suffix "*logy*", which refers to word or speech [1]. Phonology deals with the interpretation of speech sounds within and across words. In fact, phonological analysis employs three different sorts of rules [1]:

1) **phonetic rules:** sounds within words;

2) **phonemic rules:** pronunciation differences when words are spoken together

3) **prosodic rules:** stress and intonation differences across a sentence.

5.2. Morphology

The different parts of the word represent the smallest units of meaning (known as *Morphemes*). Thus, the morphology deals with the componential nature of words, which are composed of morphemes. Since the meaning of each morpheme remains the same across words, humans can break down an unknown word into its constituent morphemes in order to understand its meaning [1]. Similarly, a CL system can recognize the meaning conveyed by each morpheme in order to gain and represent the meaning. The words that cannot be divided and have a meaning by themselves are called *lexical morpheme* [1].

The words that are combined with the lexical morpheme are known as *grammatical morphemes*. Grammatical morphemes can be divided into *bound* morphemes and *derivational* morphemes [1].

5.3. Lexicography

In lexicography, humans, as well as CL systems, interpret the meaning of individual words. The assignment of a single part-of-speech (POS) tag to each word is one sort of processing that contributes to word-level understanding. Words that can act as more than one part-of-speech are allocated the most likely POS tag based on the context in which they occur in this processing [1].

Words with only one conceivable sense (or meaning) can be replaced by a semantic representation of that meaning. The type of representation depends on the semantic theory used by the CL system, because all the words share the same set of semantic primitives [1]. These simplified lexical representations allow computers to unify meaning across words, and construct sophisticated interpretations in the same way that humans do.

Lexicography may require the use of a lexicon, and the method used by a CL system will determine whether a lexicon is used or not, as well as the nature and breadth of information encoded in the lexicon [1]. Lexicons can be quite simple, containing only words and their POS. On the other hand, they can be quite complex containing information on the semantic class of the word, what arguments it takes, and the semantic limitations on these arguments. In addition, they may contain definitions of the senses in the semantic representation used in a particular system, and even the semantic field in which each sense of a polysemous word is found [1].

5.4. Syntax

Syntax concerns determining the grammatical structure of a sentence by evaluating the words inside it, where the grammar and the parser are both required for this. This level of processing produces a representation of the sentence that highlights the structural dependencies between the words [1].

There are a variety of grammars that can be used, and each of which has an impact on the parser used. Not every CL application requires a sentence full parsing. The other issues related to parse prepositional phrase attachment and conjunction scoping are no longer a barrier to those applications that rely on phrasal and clausal dependencies. In most languages, syntax conveys meaning, because order and dependency play an important role in meaning [1].

5.5. Semantics

The most of people think that the meaning is decided by the semantics, however as we can see from the definition of levels above, the meaning is determined by all the levels. Semantic processing focuses on the relationships between the word-level meanings of a sentence to discover the sentence's various meanings [1]. This level of processing can involve words semantic disambiguation with several meanings. It is similar to how syntactic disambiguation of words that can serve as multiple POS is handled at the syntactic level [1].

Semantic disambiguation allows polysemous words to have only one sense chosen and incorporated in the sentence's semantic representation. If the rest of the sentence's information is needed for disambiguation, the semantic level, not the lexical level, will be used [1]. To accomplish the disambiguation, a variety of methods can be used, some of which require information about the frequency with which each sense appears in a particular corpus of interest or in general usage. Other methods consider the local context, and still others rely on pragmatic knowledge of the document's domain [1].

5.6. Discourse

The discourse level of CL works with text units that are longer than a sentence. It does not read multi-sentence texts as a series of concatenated sentences, but each one can be read alone [1].

Discourse, on the other hand, concentrates on the whole text quality, which communicates the meaning through connecting component sentences. At this level, several forms of discourse processing can take place, and the most prevalent of which are anaphora resolution and discourse/text structure recognition [1]. Anaphora resolution is the process of replacing semantically unoccupied terms, such as pronouns, with the right entity to which they refer. The roles of sentences in the text are determined by the discourse structure recognition, which adds to the text a meaningful representation [1].

5.7. Pragmatics and dialogue

Pragmatics is concerned with the purposeful use of the language in situations, and utilizes context over and above the contents of the text for understanding [1]. The goal is to explain how extra meaning is read into texts without actually being encoded in them. This requires much world knowledge, including the understanding of intentions, plans, goals, etc. Some CL applications may utilize knowledge bases and inferencing modules.

6. Different applications

Will give a brief overview about some common CL applications that are still a timely research field.

6.1. Machine translation

Machine translation (MT) is described as a translation from one language to another carried out by software without human intervention, at least when the "raw" translation is rendered. In order to identify, research, and attempt to solve different challenges involved in automatic translation, MT systems involve advanced computational linguistics [12].

MT has numerous advantages such as increasing the productivity through automation, accessing to instantaneous results without relying on the availability of human translators and the ability to translate any text whatever the source or the length. MT systems can be divided into three distinct categories [13]:

◆ **Rules-based MT**: is based on grammatical rules combined with standard lexicons and dedicated dictionaries.

◆ **Statistical MT:** rather than focus on linguistic rules, statistical MT models undergo a learning process using a significant amount of data fed to the system to analyze (generally at the sentence level).

◆ **Neural MT**: new MT approaches are based on word sequencing, and uses sophisticated networks to process units. These neural networks imitate the human brain's neural networks, and try to predict the suitable translation.

MT is considered as one of the most difficult areas of computational linguistics, because it requires deep knowledge in different fields, i.e. computer science, linguistics, language cognition, translation, and language description methods, etc.

6.2. Information retrieval

CL could help to find a missing piece of the puzzle in unstructured data. An information retrieval system indexes a set of documents, analyzes the user's query, compares the descriptions of each document to the query, and displays the appropriate results.

6.3. Information extraction

A computational linguistics system with the goal of quickly gathering and extracting useful information for use in the growth and improvement of various enterprises. Information extraction uses powerful algorithms and software applications to convert unstructured data from social network conversations, emails, and contacts with customer care agents into accessible data that helps businesses make better decisions [14].

6.4. Chat bot

The essential asset of any company is the customer service and experience, in order to improve the manufactory quality and delivery service by ensuring the client satisfaction. Unfortunately, this task is tricky and time consuming by direct communication with all clients.

Chat-bots play a substantial role in such scenario to assist businesses by accomplishing the company goals and ensuring positive client experiences. They do not only make businesses more easier, but they also save customer from frustration to wait long time to get conversations with customer service. In addition, such system allows to save money on recruiting customer service agents.

Chatbots were formerly a merely tool for answering customers' questions, but they have since evolved into a personal companion. They may do everything from promoting a product to soliciting consumer feedback [15].

6.5. Text summarization

Text summarization is the task that consists of extracting relevant information from a text in order to create a summary automatically, which speed up the process of sifting through massive volumes of data in news articles, legal documents, etc. Generally, there are two methods of text summarization:

- **Extraction-based summarization:** takes essential words and generates a summary without adding any further information
- **Abstraction-based summarization:** paraphrases the original content to produce new terms.

6.6. Text generation

Text generation is a subfield of natural language processing. It uses computational linguistics and artificial intelligence knowledge to generate natural language texts automatically that can meet specific communication needs.

6.7. Word disambiguation

Word sense disambiguation is a challenge in NLP, which involves detecting the sense (or meaning) of a word that is activated by its use in a certain context [16]. It may appear unconscious in individuals, and it is a natural classification problem. For instance, given a word and its potential senses as defined in a dictionary, we sort all the word occurrences regarding a given context by a descending order, and we take the most relevant.

6.8. Text simplification

Text simplification is the process of decreasing the reading (and comprehension) complexity of a text, by simplifying its vocabulary and sentence structure while maintaining the original content. The purpose of such task is to increase the accessibility for people with cognitive disorders such as aphasia, dyslexia, and autism [17]. It could be also used by non-native speakers and youngsters with reading challenges.

6.9. Stemming

The stemming process consists of transforming a given word to its root or stem. Generally, it is based on removing a portion of the word in the smart way, where it can decide how to cut-off the word [18]. Reducing a word to its root or stem is a complex task, and needs a deep linguistic knowledge about the target language.

7. Text categorization approaches

Text categorization can be done in two ways: manually or automatically. A human annotator interprets the text substance and categorizes it properly and manually. This procedure can produce excellent results, but it is time-consuming and costly. AI-guided approaches are used to automatically categorize the texts in faster, economical, and more accurate manner.

7.1. Rule-based techniques

Rule-based techniques use a collection of customized language rules to classify the texts into known categories. These rules teach the system to find suitable categories based on the content of a text by using semantically relevant components of the text [19].

Rule-based systems are understandable by humans and can be enhanced over time. However, there are several drawbacks with this strategy. In particular, these systems require a deep understanding of the domain, as well as they are time-consuming, because developing rules for a complicated system can be difficult and often requires extensive research and testing. Rule-based systems are particularly challenging to maintain and scale, since adding new rules can alter the results of previously applied rules.

7.2. Machine learning

Machine learning based text categorization learns to create classifications based on past observations rather than the humanly generated rules. Machine learning algorithms may understand the varied correlations between the pieces of the text, and the specific output is expected for the specific input (using labeled training data).

Feature extraction is the primary step of training a machine learning NLP classifier. It is a heuristic employed to represent each text numerically in a form of a vector [20]. Bag of words (BOG) is one of the common used representations, in which a vector indicates the frequency of a word in a precompiled lexicon of terms. Some of the most popular text classification algorithms include the Naïve Bays family of algorithms, support vector machines (SVM), artificial neural networks (ANN), etc.

7.3. Deep Learning

Deep learning, often known as neural networks, is a set of algorithms and approaches inspired by how the human brain works. Deep learning architectures provide a lot of advantages for text classification, since they can perform extremely well with low-level engineering and computation.

Deep learning (DL) algorithms require a large amount of training data than traditional ML algorithms. However, unlike standard machine learning algorithms (i.e. SVM and NB), DL classifiers do not have a learning threshold for the training data, therefore as more data are fed as more the algorithm will be well trained.

8. Conclusion

This chapter presents a background about the artificial intelligence and an introduction to computational linguistics. In addition, we have highlighted a brief history of the computational linguistics and some challenges facing CL.

We have also discussed some fundamentals about different levels of linguistics and their role in improving CL systems. Moreover, common and challenging CL applications have been stated in this chapter with their objectives.

Finally, this chapter ends with an overview about text categorization, in which our work is classified (i.e. offensive language identification or classification). In the next chapter, we will define the offensive language and its different categories, as well as the related research works undergone in this area.

Chapter II :

Offensive

Language

1. Introduction

According to new research [21], offensive language on social media is a major problem that impacts a lot of individuals and groups. As a result, automatic systems of detecting objectionable language have been the subject of numerous studies. For the categorization of offensive texts, several statistical, machine learning, and deep learning approaches were used in the literature.

In this chapter, we give an overview about the formal langages and the dialects, as well as we define the offensive language with its categories and the depression outcomes. Furthermore, we present a literature review about previous research works carried out on automatic offensive language identification.

2. Formal languages

A language is a mean of communication between humans, and is specific to a community or country. Therefore, it is spoken and written human communication that involves the use of words in a structured and conventional manner (i.e. it has a standard orthography). This general principle can be used to any language, including sign and image-based languages [22].

2.1. Latin languages

Indo-European languages are a group of languages spoken over the Europe countries, as well as some of southwest and south Asia. Indo-European languages fall into branches, these are some of them (well-known):

- **Greek:** despite its numerous dialects, it has been a single language throughout its history. It has been spoken in Greece since at least 1600 BCE[1], and probably since the end of the 3rd millennium BCE [23].
- **Italic:** the Italic group's primary language is Latin, which is the ancestor of modern Romance languages such as Italian, Romanian, Spanish, Portuguese, French, etc. The first Latin inscriptions are thought to be from the 6th century BCE, with a literature from the 3rd century [23].
- **Germanic:** Germanic tribes lived in southern Scandinavia and northern Germany in the first millennium BCE. From the 2nd century BCE onwards, their expansions and

[1] BCE is the abbreviation of Before the Common Era

migrations are well documented in the history [23]. The Gothic of the 4th century CE is the oldest Germanic language. English, German, Dutch, Danish, Swedish, Norwegian, and Icelandic are among the other languages spoken.

The Indo-European languages are grouped together, because they share numerous basic vocabulary items, such as grammatical affixes, whose shapes in different languages can be connected to one another using statable phonetic laws [23]. The shared patterns of sound alternation are particularly essential.

2.2. Arabic language

Around 420 million people use Arabic as their native language around the world. In Arabic, the script is read and written from right to left. It has 28 letters, each of which can be written in a variety of ways depending on where it is located in the word. Vowels are represented using diacritics. There are two forms of Arabic:

- **Classical Arabic (CA):** often known as Quranic Arabic, and is the written language of the holy Quran (Islam's sacred and spiritual text). Classical Arabic is no longer a spoken language and is mostly utilized for religious purposes due to its age.
- **Modern Standard Arabic (MSA):** is the Arabic world's official language, and is the media's and culture's predominant language. It is used in formal meetings, politics, news, newspapers, etc. MSA is based on CA in syntactic, morphological, and phonological terms, but modern lexically. It is not an Arab's native language, but it is the language of instruction throughout the Arab world, but it is a written language rather than spoken one.

3. Dialects

A dialect, according to the dictionary is "*a distinctive version of a language that is peculiar to a given place or social group*". This suggests that a language can be viewed as a parent, with a variety of dialects derived from it.

3.1. Indo-European dialects

European Romance languages have many local dialects, which form a continuum of varieties that cross-country borders and stretch from Portuguese in the west to Romanian in the east [24]. Purely linguistic criteria to distinguish between a language and a dialect are hard to apply systematically, and decisions are often made on political, cultural, and literary grounds [24].

In Spanish, the traditional distinction is drawn between Spanish from Spain (Castilian Spanish) and Spanish from Latin America (Spanish from the Americas). There are differences in pronunciation, grammar, vocabulary, and intonation within each dialect [24]. Despite many geographical distinctions, Spanish speakers from various nations can communicate easily together.

The Iberian Peninsula dialects and the Brazilian dialects are the two primary groups of dialects in Portuguese. Pronunciation, grammar, and vocabulary are among the variations between both. African and Asian Portuguese dialects are more similar to those spoken in Portugal than in Brazil [24].

There are several significant dialects in French. Nevertheless, European French is commonly grouped into two primary dialects, each of which encompasses a number of regional variants. Northern and Central variations of French (language d'oil), as opposed to Southern varieties of French (langue d'oc), are the two major variants in European French [24]. In terms of pronunciation, vocabulary, and syntax, Canadian French variations differ from Standard French. Quebecois, Franco-Ontariens, and Acadiens are the three kinds of French spoken in Canada. More than 100 million people speak French as the first or second language in 31 African countries [24]. In terms of pronunciation, vocabulary, and grammar, they are all different from Standard French. They are usually divided into three groups:

1- Western, Central, and East Africa
2- Northwest Africa
3- Islands in the Indian Ocean.

Italian dialects form a continuum of intelligibility, although geographically distant varieties are not mutually intelligible. Even though standard Italian is the sole written language in modern Italy, and the majority of people speak in regional dialects [24]. Romanian is typically split into three dialects [24]:

1. Eastern Romanian (includes Moldovan);
2. Western Romanian (Transylvanian);
3. Southern Romanian (includes Muntenian/Wallachian, which has been adopted as the country's official language).

Greek dialects were split into three groups: Doric, Aeolic, and Ionic. They identified three ethnic subdivisions or "tribes" among themselves, and these groups correspond to them. The Greek dialects are divided into four sub-families by modern dialectologists, i.e. West Greek, Arcado-Cypriot, Attic-Ionic, Aeolic [24].

The English language is spoken across the world, and it was estimated that there are over 160 dialects of the language [24]. Because of the variances in delivery and pronunciation among local cultures around the world, this number is rapidly rising.

3.2. Arabic dialects

In everyday life, Arabic dialects are the common communication way used by Arabic people. Because of Arabic dialects lack a standard orthography, they might differ even within the same state of a country such as the case of Algeria [25]. Moreover, Arabic dialects are frequently used in user-generated material on social media platforms such as Twitter, Facebook, Instagram, etc. The following are some of the most common Arabic dialects:

- Egyptian Arabic is the language of Egypt and Sudan.
- Lebanon, Syria, Jordan, Palestine, and Israel are all covered by *Levantine* Arabic.
- Gulf dialectal Arabic regroups the languages spoken in Kuwait, the United Arab Emirates, Bahrain, Qatar, Saudi Arabia (with a variety of sub-dialects) and Oman.
- Morocco, Algeria, Tunisia, Mauritania, and Libya are all covered under North African Arabic (or *Maghrebi* dialectal Arabic).
- Iraqi dialectal Arabic is a hybrid of Levantine and Gulf dialects.
- Yemenite dialectal Arabic is frequently regarded as a distinct dialect.

On social media, the Arabic language may take different form, where it could be written in Arabic script or Romanized script (called *Arabizi*).

3.3. Algerian dialectal Arabic

Algeria is the largest Maghrebian countries (northest Africa), with a surface area of about 2.4 million km^2. Algerian Arabic, or the dialect spoken in Algeria, called locally as "*Daridjah*," has a complicated linguistic structure. In particular, Arabization process resulted from the adoption of the Arabic language by the native Berber population (speaking *Berber* or *Tamazight*), and the profound French colonization over 130 years. In addition, other languages, such as Turkish, Italian, and Spanish among others, had an impact on the Algerian dialectal Arabic.

Many borrowed terms are used in the Algerian dialect, and there is a lot of code-switching (i.e. using two or more languages within the same sentence). For instance, the phrase

“عندك نيفو تاع سكولة ومازلت تهدر؟” which mean “You have an elementary school level and you still talking?”. The word ‘نيفو’ comes from the French word ‘Niveau’ which means ‘Level’, and the word ‘سكولة’ comes from the Italian word ‘Scuola’ which means ‘School’, while the other words are derived from the Arabic language.

4. Offensive language

Offensive language is a text that contains some form of abusive behavior, such as activities taken with the goal of damaging, injuring, or making people angry [26]. Offensive language can take different forms, including hate speech, violent material, cyberbullying, and toxic comments. Disturbance, disrespect, harm, insult, and anger may result from this abusive behavior, disrupting the flow of dialogue.

Offensive language can be divided into several categories regarding the degree of the offense. We may find hate speech, cyber-bullying, abuse, violence and adult content among others.

4.1. Hate speech

Hate speech is defined as a text directed against a group of people with the intention of causing harm, violence, or societal instability [26]. It is defined as a language that is intended to be disparaging, humiliating, or insulting to members of a particular group or to demonstrate hostility towards them [26]. It can also be defined as using stereotypes to refer to people based on their membership in certain groups, making negative statements about minority groups, using racial and disparaging terms to cause harm. In addition, using racial and sexist slurs, supporting organizations that promote hate speech, and discriminating based on nationalities or religions [26].

4.2. Cyber-bullying

Bullying that takes place through the use of modern technology is known as cyberbullying. It can happen on social media, messaging systems, gaming platforms, and mobile phones, among other mediums. It is a pattern of behavior intended to frighten, anger, or shame individuals who are being targeted [27].

There are several different ways to bully someone online, and some people can be bullied in more than one way. Cyberbullying can take several forms, including:

- **Harassment:** is defined as sending disrespectful, impolite, or insulting messages as well as being abusive. Comments on postings, photographs, and in chat rooms that are obscene or degrading [27]. On gaming sites, being overtly offensive.

- **Denigration:** is when someone sends false, hurtful, and incorrect information about another individual. Spreading fake rumors and gossip by sharing images of someone for the goal of humiliation [27]. This can happen on any website or application.
- **Flaming:** is when someone engages in online disputes and conflicts by purposefully using extremely insulting words. They do this to elicit reactions and take pleasure in the fact that it makes someone upset [27].
- **Cyber stalking:** entails continually sending messages containing threats of damage, harassment, or frightening remarks, as well as engaging in other online behaviors that make a person fear for his or her safety [27]. Depending on what they are doing, their acts may also be criminal.

4.3. Abuse

Abusive language refers to verbal messages that include, but are not limited to, swearing, name-calling, or any other very disrespectful, rude, or harmful language that uses profanity [26]. Abusive language can also include employing harsh, insulting language, as well as utilizing or engaging physical or emotional cruelty.

4.4. Violence

There are several classes of social media violence including: crime, terrorism, human rights violation, political opinion, crisis, accidents, and conflict. The use of manipulative or coercive language that generates fear, guilt, humiliation, praise, blame, duty, obligation, or punishment is often the cause of violent communication [26].

4.5. Adult content

Adult content can take several forms, including photos, videos, and texts, in which the promotion of adult or sexual products, services, or content is expressed [26]. Posts featuring vulgar content, such as explicit and rude sexual references and pornography, are also considered adult content.

5. Depression

Depression is a widespread mental illness that affects over 264 million individuals around the world. It is marked by a chronic sadness and a lack of interest or pleasure in formerly rewarding or pleasurable pursuits [28]. It can also cause sleep and appetite disturbances, as well as fatigue and impaired focus.

Depression is a primary cause of disability worldwide, and it contributes significantly to the global illness burden. Depression's consequences can be long-lasting or recurrent, and they can have a significant impact on a person's capacity to function and live a fulfilling life [28].

Depression is caused by a complex combination of social, psychological, and biological variables. Childhood hardship, loss, and unemployment are all factors that can contribute to and accelerate the development of depression.

5.1. Outcomes

Depression may result severe outcomes, and some of them are as follows:

5.1.1. Serious health issues

Regarding the depression and pain share a similar neural route in the brain, the persons who are depressed may also suffer from common aches and pains such as headaches, backaches, stomachaches, and joint and muscle aches [29]. Depression can also increase the risk of developing other chronic conditions like heart disease, diabetes, Alzheimer's, stroke, and others.

5.1.2. Relationship trouble

Relationships can be strained by depression, resulting in lost friendships, severed ties, and breakups or divorces. That is because depression can lead to feelings of extreme loneliness and alienation, which can make any relationship difficult. Debilitating tiredness and hopelessness are common symptoms of depression, which may be highly distressing for two people in a relationship [29].

5.1.3. Suicide

Suicide is the second most common cause of death among teenagers. Mental disease, most typically depression, is a leading cause of suicide [29]. Suicidal people are overcome by terrible emotions and perceive death as the only way out, oblivious to the fact that suicide is a permanent "solution" to a temporary state [29].

5.2. Online depression

People are now more connected than ever before thanks to social media sites like Facebook, Twitter, Instagram, and others, where they may personalize their online presence, conveying a digital persona of sorts. While social media has undoubtedly resulted in many positive outcomes, it is apparent that there have been some negatives associated with it as well.

A person's mental health can be harmed by social networking. The amount of time spent on various channels, as well as the fact that people's online lives appear to be far more glamorous and fantastic than their real lives, may be part of the reason why social media causes depression [30].

Another factor is that bullying has become more prevalent. Online bullying has evolved into insults, threats, disagreements, and harassment in chat rooms and comment sections of social media and news outlets. Reading unpleasant comments online makes people feel even more depressed, and it may even cause certain depressive symptoms. It could also make people feel anxious, especially if they read comments that make them fearful or question their own decisions.

5.3. Online depression prediction

Now, social media is being used to model mental health and to better understand health risks. Quantitative techniques are increasingly being used by computer scientists to predict the presence of mental diseases and symptomatology such as depression, suicidality, and anxiety.

Researchers in computer science are predicting the presence of mood and psychosocial illnesses using behavioral and language clues from social media data. These user characteristics can be gathered from many social networking platforms.

Social networking platforms can be used as a technique for determining the severity of depression as they are a big source of data.

Questioning and behavior reports from friends and relatives were used to diagnose depressed patients in the past [31]. However, the results were not precise or qualitative. Social media, on the other hand, can be used to generate more qualitative and accurate results.

People nowadays use many social media platforms, wherein they express their inner feelings, emotions, and thoughts. It is now simple to learn about a person's everyday activities, emotions, and viewpoint. The depression level of an individual can be predicted by analyzing the user's activities (data) and applying AI heuristics.

6. Automatic offensive language detection

In this section, we present a literature review about previous works carried out in offensive language detection on different languages.

6.1. Formal languages

6.1.1. Latin languages

A statistical classifier based on sentiment analysis has been proposed in [32], wherein the authors proposed a model to detect subjectivity. The proposed model did not detect only if the phrase is subjective, but rather it can to identify and score the polarity of sentiment expressions. Another statistical classifier has been proposed by Santucci, for which the authors collected over 1M Tweets to detect hate speech in Italian language [33]. The classifier is mainly based on FastText [34] and a heuristic of feature selection.

A machine learning approach based on feature selection of metadatahas been proposed to deal with automatic offensive language detection in Twitter [35]. In particular, the authors used SVM and NB classifier, where the latter outperformed the SVM (i.e. 92% of accuracy in contrast to 90%).

Another classifier combining SVM and MLP (Multi layer perception) has been proposed in [36], where the authors used Stochastic Gradient Descent (SGD) for feature selection. They experimented with three in Indo-European Languages (i.e. English, German and Hindi), and the results showed that the proposed framework is suitable with the Hindi language rather than English and German.

Greek offensive language identification has been proposed, wherein the authors created a new Twitter corpus (OGTD) containing 4.7K texts manually annotated [37]. The authors tested different ML and DL approaches such as SVM, SGD, NB and LSTM. The experimental results on OGTD showed that LSTM with Attention is proming and outperforms conventional ML approaches (i.e. 0.89 of F1-score).

A hate speech detection approach established a lexical baseline for discriminating between hate speech and profanity on a standard dataset [38]. Another approach presented the development of a system that can automatically detect profiles and persons that extensively and deliberately post messages containing offensive language [39].

Badjatiya et al. (2017) have annotated a Twitter corpus of 16K text for hate speech identification, and experimented with various DL approaches [40]. Another work used DL approaches has been proposed for hate speech in Indonesian Language [41]. The authors tested different feature models, where textual features produced promising results (87.98% of F1-score).

6.1.2. Arabic language

An approach in Abusive Language Detection on Arabic Social Media has been proposed, where two datasets were introduced [42]. The first contains 1,100 manually labeled tweets, and the second contains 32K user comments that the moderators of a popular Arabic news site deemed inappropriate. The authors proposed a statistical approach base on a list of of abcene words, and the produced results were around 60% of F1-score.

Detecting cyberbullying in Arabic content has been carried out in [43], wherein the authors introduced an approache focused on preventing cyberbullying attacks. The approach uses NLP to identify and process Arabic words, and ML techniques to classify bullying content.

A hate speech Arabic dataset with 9.3k annotated tweets was proposed by Raghad and Al-Khalifa [44]. The authors experimented with several DL and ML models to detect hate speech in Arabic tweets. The results showed that CNN GRU produced the best performances (0.79 of F1-score). Multitask approach covered Arabic Offensive Language and Hate-Speech Detection using DL, transfer learning and multitask learning was proposed in [45].

6.2. Dialectal languages

6.2.1. Latin languages

Offensive language detection and hate speech detection on Portuguese dialects was proposed in [46], where the authors worked on two Portuguese dialects (i.e. Brazil and Portugal). They used an offensive lexicon of implicit and explicit offensive and swearing expressions.

6.2.2. Arabic language

Husain has used different ML approaches and an ensemble classifier to deal with the offensive language identification of dialectal Arabic [47]. The conducted study showed an interesting impact of the preprocessing on such task, as well as the performance of the ensemble classifier in contrast to single ML algorithms.

Another work focused on Tunisian hate speech and abusive speech was proposed, in order to create a benchmarked dataset (i.e. 6k of tweets) of online Tunisian toxic contents [48]. The authors tested two ML approaches (i.e. NB and SVM), where the NB classifier outperformed the SVM (92.9% of accuracy in binary classification).

Hate speech detection against women in Algerian community was addressed in [49], for which the authors used word2vec and FastText with different features. From the experimental results, FastText produced a promising result in contrast to word2vec.

7. Conclusion

In this chapter, we have presented a literature review about previous works on automatic offensive language detection, where we have highlighted some works undergone on Latin languages and Arabic language. In addition, we have focused on some works carried out on dialectal languages, and the dialectal Arabic particularly.

In addition, we have discussed and defined the offensive language on social media, its categories and its social impacts on individuals. In the next chapter, we will present the adopted methodology to automatic identify offensive language on Algerian dialectal Arabic.

Chapter III :

Methodology

1. Introduction

Text classification is becoming an increasingly important the more use online platforms. Such task can be done using automatic classification or manual classification. In this chapter, we describe the adopted classifiers used in automatic classification, as well as our approach which presents a rule- based classifier.

2. State-of-the-art algorithms

Text classification algorithms are used in a wide range of software systems that analyze large amounts of textual data. The choice between classification models is determined by the type of data and the type of issue to be solved. In fact, it is usually a good idea to evaluate several classification models on a given dataset, taking into account the prediction accuracy and the processing efficiency.

2.1. Convolutional neural networks (CNN)

The conventional neural network (CNN) is a deep learning model that is designed to learn spatial hierarchies of features from low- to high-level patterns automatically and adaptively. CNNs are neural networks with one or more convolutional layers that are primarily used for image processing, classification, segmentation, and other auto-correlated data. They have also been used in Natural Language Processing and speech recognition.

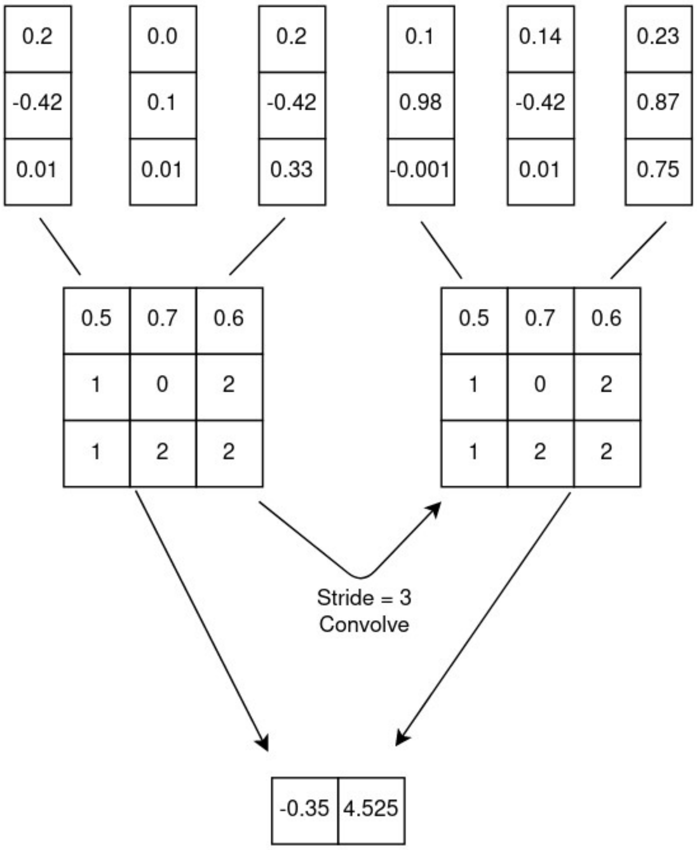


Figure 3.1. General concept of the convolution process.

The input data is represented by CNN as multidimensional arrays. The receptive field is the extraction of each part of input data by CNN. It gives weights to each neuron depending on the receptive field's importance, and it can distinguish between the significance of neurons.

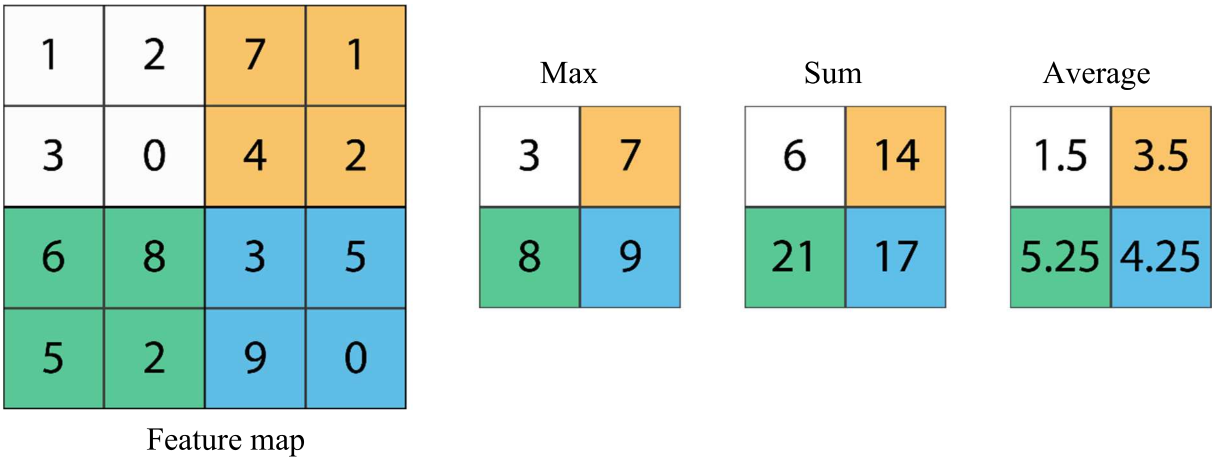


Figure 3.2. Different pooling operations

Convolution and pooling are two essential procedures that are always included in CNN. The convolution process with several filters is capable of extracting features from the dataset while preserving their spatial information. The former is a linear operation in which a collection of weights is multiplied by the input. The multiplication is done between an array of input data and a two-dimensional array of weights, called a filter or a kernel [50].

The filter is smaller than the input data, and the scalar product is used to multiply a filter-sized patch of the input with the filter. It is intentional to use a filter that is smaller than the input, because this allows the same filter to be multiplied by the input array several times at different positions in the input. From left to right, top to bottom, the filter is applied systematically to each overlapping section or filter-sized patch of the input data [50].

A single value is obtained by multiplying the filter with the input array once. The result of applying the filter to the input array several times is a two-dimensional array of output values that indicate input filtering; this two-dimensional output array is referred to as a "feature map" [50].

Pooling, also known as subsampling, and it is a technique for reducing the dimensionality of feature maps created by the convolution procedure. The pooling layer works on each feature map individually to build a new set of pooled feature maps. Pooling is similar to applying a filter to feature maps in that it involves selecting a pooling process. The pooling process or filter has a smaller dimension than the feature map. There are several types of pooling operation, max pooling takes the largest element, sum pooling extracts the sum of all elements and average pooling calculates the average value from each patch of the feature map [50].

2.2. Bidirectional long short-term memory (BiLSTM)

A Bidirectional LSTM is a sequence processing model that consists of two LSTMs, one forward and one backward. LSTM networks are a type of recurrent neural network that can learn order dependence in sequence prediction problems. This is a behavior necessary in complicated problem areas such as machine translation, speech recognition, and others.

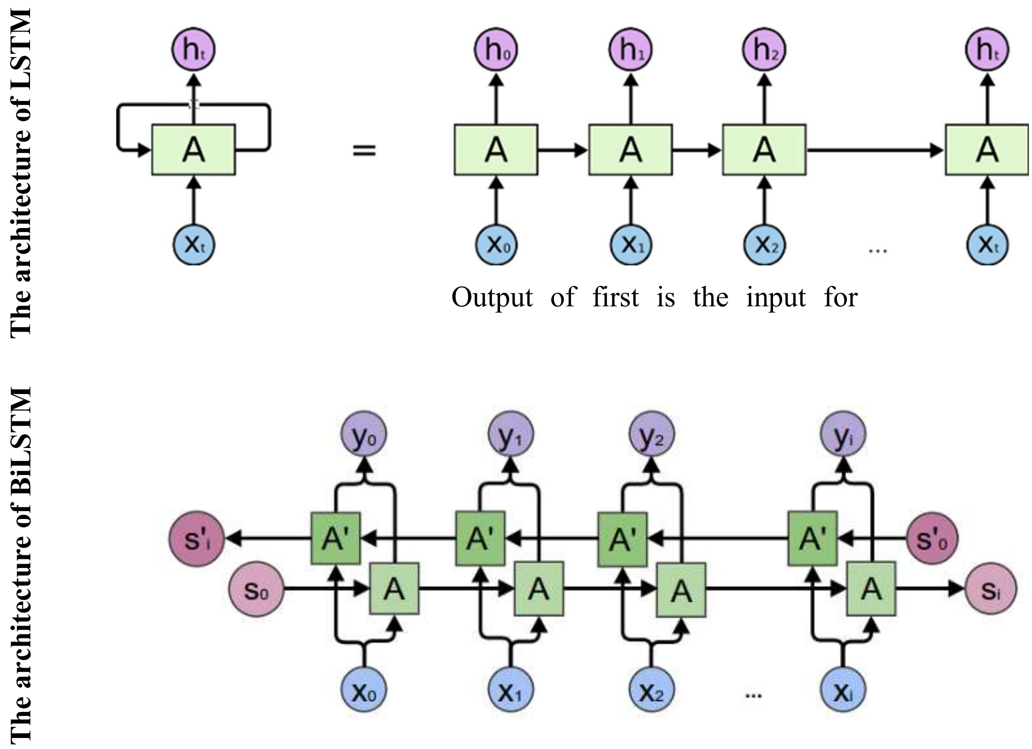


Figure 3.3. Difference between LSTM and BiLSTM

LSTMs are expressly designed to prevent the long-term reliance problem, and remembering knowledge over extended periods of time is basically their natural tendency. BiLSTM effectively improves the quantity of information available to the network, and the context provided to the algorithm [51].

What differs BiLSTM from LSTM is that it preserves information from the future and using the two hidden states combined. It is able in any point in time to preserve information from both past and future (Figure 3.3).

2.3. FastText

FastTesxt is a library intended to help with the development of scalable solutions for text representation and classification. It integrates some of the most effective techniques introduced in the last couple decades by the natural language processing and machine learning fields. These include employing a bag of words and a bag of n-grams to represent sentences, as well as leveraging subword information and exchanging information between classes via a hidden representation [52].

Figure 3.4 presents a simple illustration of the architecture of FastText. While deep neural networks reach excellent performance, they can be time-consuming to train the models. FastText can assist in resolving this issue. To be efficient on datasets with a large number of categories, it employs a hierarchical classifier rather than a flat structure, with the various categories structured in a tree (think binary tree instead of list). With regard to the number of classes, this reduces the time complexities of training and testing text classifiers from linear to logarithmic [52].

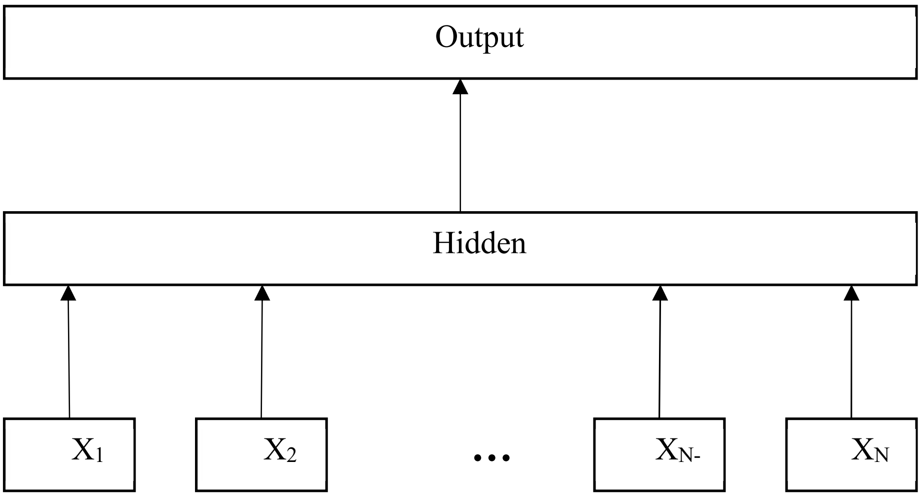


Figure 3.4. Model architecture of FastText

FastText additionally takes advantage of the fact that classes are imbalanced by employing the *Huffman* technique to construct the tree representing the categories. As a result, the depth of the tree for very frequent categories is smaller than for unfrequent ones, resulting in increased computational efficiency [52].

FastText also represents a text with a low-dimensional vector formed by summing vectors corresponding to the words in the text, where each word in the vocabulary is assigned to a low-dimensional vector in FastText. This hidden representation is shared by all the classifiers for various categories, allowing information about words learnt for one category to be used by other categories [52]. This type of representation, known as a bag of words, disregards word order. Vectors are also used in FastText to represent word n-grams in order to account for local word order, which is crucial for many text classifications issues [52].

2.4. Support vector machine (SVM)

SVM is the most frequently used machine learning (ML) based pattern classification algorithm nowadays. It was created by *Vapnik* in 1995, and is based on statistical learning theory concept of decision planes that define decision boundaries [53].

As illustrated in Figure 3.5, a decision plane is suitable for separating objects with various class memberships. The dividing line creates a barrier, where in the right side all items are GREEN, and in the left side all the objects are RED. Any new object that falls into the right of the dividing line is classified as GREEN, while anything that falls into the left is classified as RED.

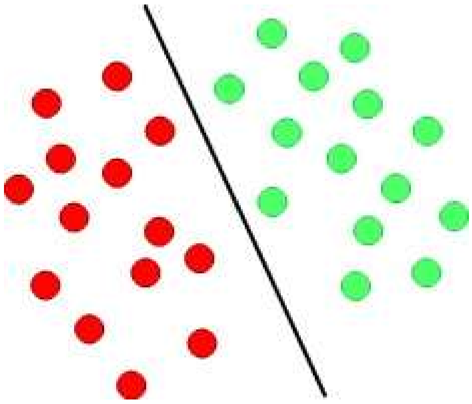


Figure 3.5. Linear separation in input data space

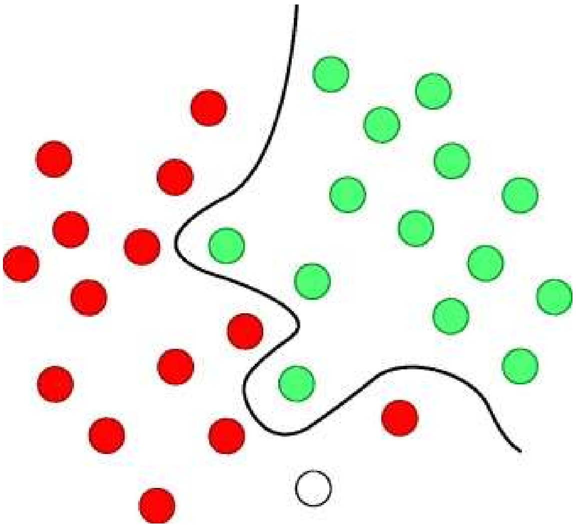


Figure 3.6. Nonlinear separation in input data space.

Most classification jobs, on the other hand, are not that straightforward, and more sophisticated structures are frequently required to achieve the best results. A complete separation of the GREEN and RED items, as shown in Figure 3.6, would necessitate a curve, which is more complicated than a line. Hyperplane classifiers perform classification problems by drawing the separation lines to distinguish between items belonging to various classes [53].

SVMs are a supervised learning method that is used for classification and regression. They are part of the generalized linear classifier family. SVM is a classification and regression prediction tool that employs machine learning theory to enhance predictive accuracy, while automatically avoiding over-fitting to the data [53]. SVMs are systems using the hypothesis space of a linear function in a high dimensional feature space, and are trained with an optimization theory learning algorithm that applies a statistical learning theory learning bias.

2.5. Naïve Bayes classifier (NB)

Naïve Bayes (NB) classifiers are linear classifiers, which are recognized for their simplicity while being extremely effective. NB is based on Bayes Theorem, which was proposed by *Reverend Thomas Bayes* in the 1760's, and the word naïve refers to the assumption that the characteristics in a dataset are mutually independent [54]. The independence assumption is frequently broken in practice, and NB classifiers nevertheless perform admirably under this unrealistic assumption.

They are used in a variety of disciplines because of their robustness, ease of implementation, speed, and accuracy. Some examples include text classification, spam filtering, sentiment analysis, real-time prediction, multi-class prediction and recommendation systems. There are several variants of NB classifier [54]:

- **Optimal Naive Bayes:** This classifier selects the class with the highest posterior probability of occurrences. It is ideal as the name implies, but running through all available possibilities is fairly slow and time-consuming.
- **Gaussian Naive Bayes:** Gaussian Bayes is based on the normal distribution. It greatly speeds up the search, and the error is just two times higher than in Optimal Bayes under some non-strict circumstances.
- **Multinomial Naive Bayes:** It is typically used to solve document classification challenges. It based its decisions on discrete criteria, such as the frequency of words in the document.
- **Bernoulli Naive Bayes:** In Bernoulli the predictors are Boolean variables. As a result, the parameters used to predict the class variable can only have yes or no values, such as whether a word exists or not in the text.

3. Our approach

Rule-based classifiers are a form of classification algorithms that make a decision based on a set of “if/else” rules. In our approach, we have created a rule-based classifier to classify a given text into three classes i.e. Offensive, Abusive or Normal.

3.1 Motivation

Recently, research in offensive language detection in Arabic Dialects has expanded greatly. However, any of these approaches addressed the Algerian dialect, and especially creating corpora for this dialect, where it is important to create a new corpus. In fact, with the growth of the number of users in social media platforms, and with offensive content left unattended, people’s mental and physical health is at risk. Thus, research in this topic usually use automatic classification, and experimenting with Algerian Dialect requires a rule-based algorithm, because of the complexity and diversity of the dialect grammar and vocabulary.

3.2 Rule based algorithm

The proposed algorithm is based on two steps, i.e. preprocessing and identification rules. The latters are manually designed by predicting the offensive and abusive writing style:

3.2.1 Preprocessing

We have applied some preprocessing steps before executing our algorithm. We have normalized some letters by remplacing them with others such as:

- replacing “hmazated alif (‘إ‘ ;’آ’,’أ’)” with “alif bare (‘ا’)”

- replacing “alif maqsura (‘ى’)” with “yaa (‘ي’)”

We have also removed punctuation marks like exclamation marks, question marks, commas and full stops.

3.2.2 Identification algorithm

We used Python to build the identification algorithm, because of its simple syntax and the number of open-source libraries available. First, we have imported some python libraries to help us preparing our corpus and achieve better results, then we have imported our offensive, abusive and normal datasets using io and json libraries. Subsequently, we have implemented the preprocessing rules to normalize the text. The algorithm relies on some rules to determine whether the text if offensive, abusive or normal, those rules are as follows:

```
For comment in text
    For abuse in absudecitonnary
        Comment  → Abusive  If  abuse ∈ comment.
        If abuse ∉ comment
            For word in comment
                Comment  → Abusive  If  word ≈ abuse #with 0.90 similarity threshold.
                If word ≠ abuse
              For pronoun in pronoundictionary
                  Comment → Abusive  If  word ≈ abuse and word(-1)= pronoun.
                  #with 0.70 similarity threshold
    If  CountVect = PredVect
        For offense in offendictionairy
            Comment  → Offensive  If  offense ∈ comment.
            If offense ∉ comment
                For word in comment
                    Comment  → Offensive  If  word ≈ offense #with 0.90 similarity threshold.
                    If word ≠ offense
                     For pronoun in pronoundictionary
                       Comment → offensive  If  word ≈ offense and word(-1)= pronoun
                       #with 0.70 similarity threshold
     If  CountVect = PredVect
              Comment → Normal.
```

Figure 3.7. Rules based offensive and abusive language detection

4. Conclusion

In this chapter, we have presented the general scheme of some state-of-the-art ML classifiers i.e. SVM and NB, and DL classifiers i.e. CNN, BiLSTM and FastText. These algorithms are commonly used in different text categorization tasks, and produced promising results. Besides these algorithms, we have presented our approach that relies on some fixed rules, where the latters were manually designed by analyzing different writing styles.

Chapter IV: Results and experiments

1. Introduction

In this chapter, we present the details about the corpus collection and annotation and some statistics. Moreover, we plot and discuss different results obtained by the adopted classifiers, where we present the results of the binary classification (offensive and normal) and the three-class classification (offensive, abusive and normal).

2. Corpus description

Offensive language detection is one of the challenging and interesting topics in research. Because of the lack of research and resources in dialects and under-resourced languages, building a corpus is challenging. We have built a new corpus for Algerian offensive language detection, where we have collected data from Facebook and manually annotated the corpus into three labels, i.e. offensive, abusive and normal.

2.1. Data collection

Algerian dialectal Arabic ("*Darija*") is the main communication language used in social media networks in Algeria. Among the social media networks, Facebook is the first ranked one and used by the most of the Algerian community whatever the age bracket [25]. In this regard, we have Facebook as a source of information to create our corpus.

Firstly, we have selected a set of public pages and groups related to politics, joking and sports subjects. Subsequently, among the page publications, we have retrieved those tackling a harmful and provoking subject that involves more interactions and comments (disliked by the majority). In particular, the subjects may contain conflicts or controversy like football, news, ethnicity and religion, because the Algerian community is conservative and some topics involves aggression while expressing opinions or judging people. Moreover, we notice that the most of the publications contain hate or abuse towards an individual, group or entity.

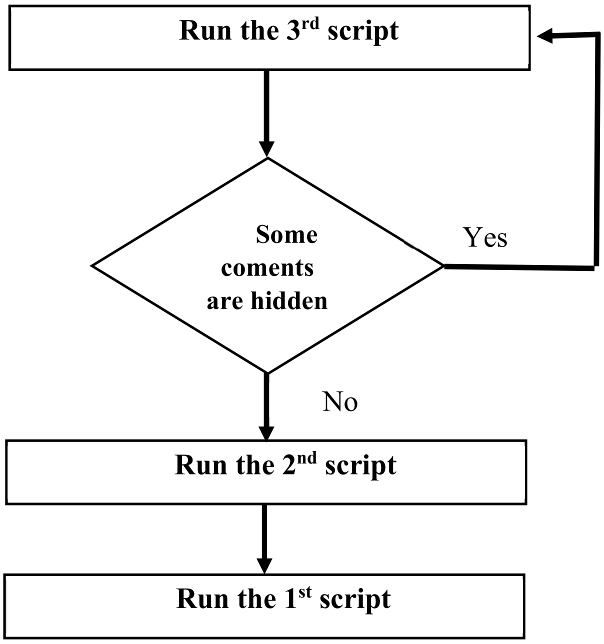


Figure 4.1. General scheme followed to crawl Facebook comments

Since Facebook new politics restricts automatic gathering data from this platform, we have manually collected the user comments. More specifically, we have created a Javascript[1] scraping tool to collect data automatically instead of standard copy/paste process, because the latter may take a very long time when collecting a large amount of data. The script analyzes the whole HTML code of the page, and extracts the comments with some relative information. Besides the comment text, the script also crawls the user's name, profile link, number of reactions, number of replies.

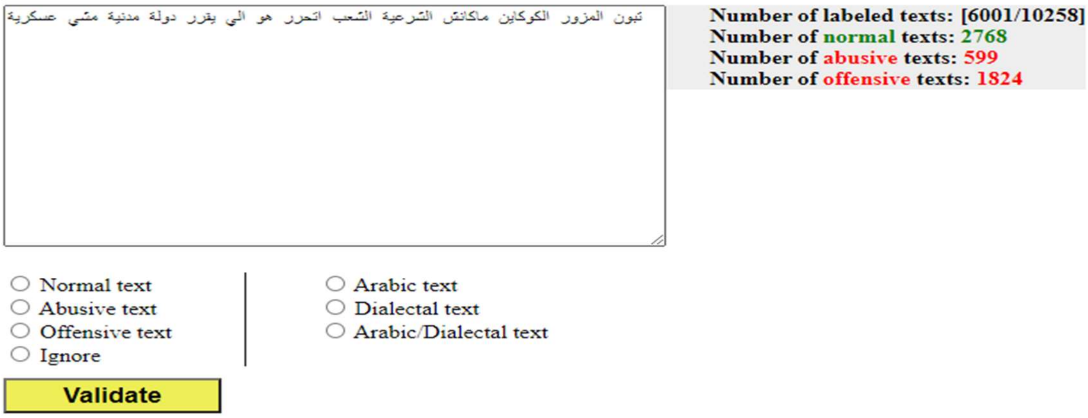
We have created other scripts to unhide the second part of long comments. It is worth mentioning that by default Facebook shows the first words of long comments and hides the others for better readability, and to optimize the bandwidth as well. In addition, we have created another script to view all the comments related to a given publication, because Facebook shows by default the most relevant comments (e.g. 30 comments) and hides the other to also optimize the navigation bandwidth.

Therefore, in the browser console (development console of each browser), we run the third script (recursively) to view all the comments by automatically clicking on the button "*View more comments*". Next, we run the second script to unhide the second part of the long comments by automatically clicking on the button "*Read more...*". The scripts work on computer browsers, and compatible with smartphone browsers.

Empty comments and comments with only images and emojis are ignored. In overall, we have crawled 10,258 comments, where each publication comments are saved in Json file containing the publication ID.

2.2. Data annotation

The corpus annotation is carried out via an in-house. In order to create a standard corpus following the scientific methodology, five annotators, i.e. Algerian native speakers, have been involved to annotate the corpus. They were unstructured to attribute one label corresponding to the nature of the text, i.e. offensive, abusive or normal, as well as one label corresponding the text language, i.e. Arabic (Modern standard Arabic), dialect or a mixture of Arabic and dialect (Figure 4.2).



تبون المزور الكوكاين ماكانش الشرعية الشعب اتحرر هو الي يقرر دولة مدنية مشي عسكرية

Number of labeled texts: [6001/10258]
Number of normal texts: 2768
Number of abusive texts: 599
Number of offensive texts: 1824

○ Normal text
○ Abusive text
○ Offensive text
○ Ignore

○ Arabic text
○ Dialectal text
○ Arabic/Dialectal text

Figure 4.2. Screenshot of the crowdsourcing platform of the annotation

[1] https://github.com/xprogramer/fb-cmt-crawl

The annotators were instructed to ignore all the texts that are completely written in French, English and Berber. In addition, unclear texts and difficult texts to label (for the offensive task) were ignored. If the text contains any kind of hate, aggressiveness, bullying, harassment or violence, or it offends someone, a group of people or an entity, it is labelled as "*offensive*". On the other hand, the text is labeled as "abusive" if it contains swear words and sexual/adult content.

Table 4.1. Statistics of the annotation rounds

	First round (Nb. texts)	Second round (Nb. texts)	Third round (Nb. texts)
Annotator #1	6,000		4,258
Annotator #2	4,258		6,000
Annotator #3	0	3,000	
Annotator #4	0	3,000	
Annotator #5	0	4,258	

Table 4.2. Corpus description

Category	Number of texts
Offensive	3,227
Abusive	1,334
Normal	4,188
Ignored	1,509
Total	*10,258*

The annotation is performed in three rounds. In the first one, annotator #1 annotated the first 6,000 texts, while annotator #2 annotated the remaining 4,258 texts. In the second round, annotator #3 annotated the first 3,000 texts, annotator #4 annotated the second 3,000 texts and annotator #5 annotated the remaining 4,258 texts.

If there is a conflict in the text labels between the two rounds, a third round is performed to attribute another label. In this regard, annotator #2 annotated the first 6,000 texts (only texts with conflicts), whereas annotator #1 annotated the remaining 4,258 texts (only texts with conflict). We applied the majority voting to attribute the final labels, where among the 10,258 texts, 3,227 texts were labelled as offensive, 1,334 texts were labelled as abusive and 4,188 texts were labelled as normal. The remained texts were ignored for different reasons as stated above.

3. Setup and configuration

After the corpus collection and annotation, the final data was organized and save in JSon files (i.e. each file for each category). We have carried out some configurations before experimenting with machine learning and deep learning classifiers in Python[2].

We have used Keras to implement deep learning classifiers and Scikit-learn to implement machine learning classifiers. On the other hand, we have written completely the algorithm of our rule-based identification. Some additional Python libraries have been used to read and prepare data such as "json" and "io". For automatic classification, and after the text is preprocessed, some tools were used to prepare the data for training.

A. Tokenization

Every text was tokenized into a vector of words (sequence of tokens) using Keras tokenizer tool (Tokenizer object). Subsequently, the sequences were padded to the same size (maximal text length) with Keras Pad_sequences tool. The latter adds a sequence of '0' at the end of the vector until reaching the maximal length.

B. TF-IDF

For ML models we used Tf-Idf Transformer tool to scale down the impact of words that occur very frequently in both normal and offensive categories, which are less informative then words that only occur in a specific category, then we trained our models using SVC, MultinomialNB and GaussianNB classifiers.

C. Corpus splitting

At the end, the corpus was divided into two sets, i.e. training and test set. The first one constitutes 90% of the whole size, while the second is 10%. For deep learning classifiers, the training set was also divided into two sets, i.e. training set with 90% and validation set with 10%.

4. Experimental results

We have conducted two sets of experiments. In the first set, we have merged the offensive and abusive categories into the same category, i.e. offensive category. Thus, we have conducted a binary classification (offensive and normal) to test the reliability of the classifiers. In the second set of experiments, we did a three-labels classification to study how well the classifiers can deal with multi-labels.

Table 4.3 presents the results of the three-labels classification obtained by the machine learning classifiers, while Table 4.4 presents the results of the binary classification obtained by the machine learning classifiers. Table 4.4 presents the results of the three-labels classification obtained by the deep learning classifiers, while Table 4.5 presents the results of the binary classification obtained by the dep learning classifiers. Table 4.7 presents the results of the three-labels classification obtained by the rule-based identification algorithm, while Table 4.8 presents the results of the binary classification obtained by the rule-based identification algorithm.

[2] Python is programming language

Table 4.3. Test results of ML models trained on three-labels

Classifiers		Category				
			Precision	Recall	F1-score	Accuracy
SVM		Offensive	0.66	0.57	0.62	**0.69**
		Abusive	0.64	0.48	0.55	
		Normal	0.71	0.83	0.77	
NB	Multinomial	Offensive	0.67	0.57	0.62	0.70
		Abusive	0.98	0.33	0.49	
		Normal	0.68	0.91	0.78	
	Gaussian	Offensive	0.56	0.49	0.52	0.53
		Abusive	0.28	0.55	0.37	
		Normal	0.72	0.57	0.63	

Table 4.4. Test results of ML models trained on two-labels

Classifiers						
			Precision	Recall	F1-score	Accuracy
SVM		Offensive	0.74	0.82	0.78	**0.75**
		Normal	0.77	0.67	0.72	
NB	Multinomial	Offensive	0.74	0.85	0.79	**0.77**
		Normal	0.80	0.67	0.73	
	Gaussian	Offensive	0.76	0.65	0.70	0.71
		Normal	0.66	0.77	0.71	

Table 4.5. Test results of DL models trained on three-labels

Classifiers					
		Precision	Recall	F1-score	Accuracy
CNN	Offensive	0.35	0.95	0.52	0.35
	Abusive	0.06	0.01	0.01	
	Normal	0.33	0.01	0.02	
BiLSTM	Offensive	0.35	0.37	0.36	0.36
	Abusive	0.16	0.13	0.14	
	Normal	0.44	0.45	0.45	
FastText	Offensive	0.61	0.56	0.58	**0.65**
	Abusive	0.43	0.97	0.58	
	Normal	0.75	0.68	0.71	

Table 4.6. Test results of DL models trained on two-labels

Classifiers					
		Precision	Recall	F1-score	Accuracy
CNN	Offensive	0.52	1.00	0.69	0.52
	Normal	0.33	0.00	0.00	
BiLSTM	Offensive	0.51	0.41	0.45	0.50
	Normal	0.50	0.60	0.55	
FastText	Offensive	0.73	0.73	0.73	**0.72**
	Normal	0.70	0.70	0.70	

Table 4.7. Test results of rule-based algorithm with three-labels.

Classifier		**Precision**	**Recall**	**F1-score**	**Accuracy**
Rule-based algorithm	Offensive	0.78	0.30	0.43	0.65
	Abusive	0.82	0.63	0.71	
	Normal	0.60	0.93	0.73	

Table 4.8. Test results of rule-based algorithm with two-labels.

Classifier		**Precision**	**Recall**	**F1-score**	**Accuracy**
Rule-based algorithm	Offensive	0.87	0.43	0.57	0.67
	Normal	0.60	0.93	0.73	

Binary classification experiments showed better overall results compared with experiments with 3 classes, and that is because more the number of classes increases the more it is complicated to distinguish between them. In addition, offensive and abusive text can be very similar sometimes which may feed conflicted information to the model which in turn causes a drop in accuracy and F1-score.

Machine Learning models showed high prediction results, where the SVM and Multinomial NB obtained the highest score, with an accuracy of 0.75 and 0.77, and 0.79 F1-score for offensive text and 0.73 for normal text. While GaussianNB achieved worst results with an accuracy of 0.71 and F1-score of 0.70 for offensive and 0.71 for normal.

Deep Learning models showed low performance, where CNN produced a high F1-score of 0.69 in the offensive class, but unfortunately the normal class F1-score was zero. In addition, BiLSTM achieved the lowest score with an accuracy of 0.50 and F1-score of 0.45 for offensive and 0.55 for normal. Conversely, FastText outperformed both CNN and BiLSTM with and accuracy of 0.72 and F1-score of 0.73 for offensive and 0.70 for normal, but it still preformed less than SVM and MultinomialNB.

Rule-based algorithm showed a good performance compared to deep leaning models, with an accuracy of 0.67, which outperformed CNN and BiLSTM, but it has still a lower performance than FastText, SVM and MultinomialNB.

In overall, DL models need a huge training set in order to achieve high performances. Nevertheless, the reason behind the CNN that produced 0.00 score on recall and F1-score for normal class, is mainly because CNN could not differentiate between normal and offensive text because of the stop words and frequent words, as they occur frequently in offensive texts. Therefore, CNN considered them as offensive words, which led to misclassifying all normal texts as offensive because of these words.

Rule-based algorithm produced a descent performance, but it is still low compared to the results of SVM and MultinomialNB, and that is because the created rules couldn't cover all the possibilities. However, it can be improved by adding rules that overcome the misclassified texts. It is obvious that the rule-based algorithm can outperform all automatic classifiers, because it uses specific rules for the specific language that is studied. These rules are designed based on the human knowledge of the language vocabulary, and it is easy to improve and expand on larger datasets.

Finally, the best overall score is produced by SVM and MultinomialNB. The main reason ML models were able to get such high results is because common words and stop words that are used in all categories were given low frequency using Tf-Idf, while unique words were given high frequencies.

5. Conclusion

In this chapter, we have presented a new corpus created for the offensive content detection, where it concerns the Algerian dialect. The corpus was crawled from Facebook social media, and manually annotated by five different annotators in three rounds. The final labels were chosen based on the majority voting. In overall, the final corpus contains 8.7k texts among 10.2k collected texts, where the remaining were ignored for various reasons.

We have evaluated the state-of-the-art algorithms (machine learning and deep learning) on this corpus, as well as our proposed algorithm. The experimental results carried out on two sets of configurations (binary classification and three-label classification) showed that our proposal produced acceptable results, but lower than some machine learning algorithms. In addition, SVM and Multinomial NB outperformed all the other algorithms because of the Tf-Idf technique that gives high weights to specific terms and low weights to common words.

Finally, our proposal could be extended by incorporating new rules to cover more writing possibilities, but it requires a deep linguistic study.

General Conclusion

General Conclusion

In this manuscript, we have presented our approach to deal with offensive language detection, where we have focused on the Algerian dialect language. The tackled problem is a big step towards protecting individuals and groups from the harm and dangers inflicted by offensive content in social media. Our approach consists of building models that are capable of classifying text into offensive, abusive and normal.

We have studied the fundamentals of computational linguistics and text categorization methods, and we have highlighted different types of languages and to distinguish Algerian Dialect from Modern Arabic. Subsequently, we have defined the offensive language with it categories, risks and outcomes. We have briefly discussed different machine learning and deep learning algorithms that are commonly used in text categorization.

In this work, we have crawled a corpus of 10,258 comments from Facebook social media. The texts were annotated at the sentence level for the offensive detection task by five annotators in three rounds. The final corpus contains 3,227 offensive, 1,334 abusive and 4,188 normal texts. The corpus was also annotated at the sentence level for the language detection, i.e. Arabic, dialect or mixed between Arabic and dialect. In addition, we have proposed a new algorithm based on rules to detect offensive and abusive content, for which we have conceived a lexicon of offensive words with 606 words and another for abusive words with 74 words.

To create our models, we created an Algerian Dialect corpus, by collecting comments from Facebook posts. We preprocessed the collected dataset to prepare it for training and testing. We used several classification algorithms to train our models with the created corpus. We have evaluated six automatic classification models on our corpus to assess the performances of the state-of-the-art algorithms on such textual data. In particular, we have conducted two sets of experiments, where the first one is a binary classification (merge offensive and abusive texts), and the second one is three-categories classification (eacho category is taken independently). We have also evaluated and compared our proposal with the state-of-the-art algorithms, where the results showed interesting performances, but the algorithm requires further investigation to fine tune the rules of the lexicon. In overall, SVM and NB keep the best identification scores and outperform deep learning classifiers, because the latters require a huge training set to well extract the features, especially in the case on unstructured texts.

In future work, our corpus needs to be expanded to cover more offensive content, and our rule-based algorithm needs to be investigated to solve language complexity problems and to achieve better classification results.

Bibliography

Bibliography

1. Ruslan Mitkov: The Oxford Handbook of Computational Linguistics, Oxford university press, 786 pages, (2004).
2. Joost Nico Kok: ARTIFICIAL INTELLIGENCE, EOLSS Publications, 418 pages, (2009).
3. Ethem Alpaydin: Introduction to Machine Learning, second edition, MIT Press, 584 pages, (2009).
4. Ralph Grishman: Computational Linguistics: An Introduction, Cambridge University Press, 193 pages, (1986).
5. Igor A. Bolshakov and Alexander Gelbukh: Computational Linguistics: Models, Resources, Applications. Instituto Politécnico Nacional, 186 pages, (2004)
6. Jacob Eisenstein: Introduction to Natural Language Processing, MIT Press, 536 pages, (2019).
7. James Allen: Natural language understanding, Benjamin-Cummings Publishing Co, (1988).
8. Stefano Ceri, Alessandro Bozzon, Marco Brambilla, Emanuele Della Valle, Piero Fraternali, Silvia Quarteroni: Web Information Retrieval. Springer Science & Business Media. 284 pages. (2013).
9. Karen Sparck Jones and Prter Willett: Readings in Information Retrieval, Morgan Kaufmann, 589 pages, (1997).
10. D.I. Hernández Farias, P. Rosso: Chapter 7 - Irony, Sarcasm, and Sentiment Analysis, Morgan Kaufmann, Pages 113-128, (2017).
11. Katz, Graham and Diab, Mona: Introduction to the Special Issue on Arabic Computational Linguistics, Association for Computing Machinery, (2011)
12. M. D. Okpor, Machine Translation Approaches: Issues and Challenges, International Journal of Computer Science Issues, (2014)
13. Sneha Tripathi and Juran Krishna Sarkhel: Approaches to machine translation, NISCAIR-CSIR, pages 388-393, (2010).
14. Sunita Sarawagi; Information Extraction, Now Publishers, 132 pages, (2008)
15. Tarun Lalwani, Shashank Bhalotia, Ashish Pal, Shreya Bisen, Vasundhara Rathod: Implementation of a Chatbot System using AI and NLP. International Journal of Innovative Research in Computer Science & Technology, 5 Pages, (2018).
16. Edilson A. Correa Jr., Alneu de Andrade Lopes and Diego R. Amancio: Word sense disambiguation, Information Sciences 442-443, pages 103-113, (2018).
17. Advaith Siddharthan: A survey of research on text simplification. ITL - International Journal of Applied Linguistics, pages 259 – 298. (2014).
18. K. Abainia, S. Ouamour and H. Sayoud: A Novel Robust Arabic Light Stemmer. Journal of Experimental & Theoretical Artificial Intelligence (JETAI'17), pages 557-573, (2017).
19. Tung A.K.H: Rule-based Classification. Encyclopedia of Database Systems, (2009)
20. D. Michie , D. J. Spiegelhalter , C.C. Taylor: Machine Learning, Neural and Statistical Classification. DOI:10.2307/1269742, (1999).
21. Naman Deep Srivastava, Yashvardhan Sharma: Combating Online Hate: A Comparative Study on Identification of Hate Speech and Offensive Content in Social

Media Text. IEEE Recent Advances in Intelligent Computational Systems, Pages 47-52, (2020).

22. Victoria Fromkin, Robert Rodman, Nina Hyams: An Introduction to Language, Cengage Learning, 624 pages, (2018).
23. Mate Kapović, Anna Giacalone Ramat, Paolo Ramat: The Indo-European Languages, 650 pages, (2017).
24. Philip Baldi: An Introduction to the Indo-European Languages. SIU Press, 214 pages, (1983).
25. Abainia, K. DZDC12: a new multipurpose parallel Algerian Arabizi–French code-switched corpus. Lang Resources & Evaluation 54, 419–455 (2020).
26. Husain, Fatemah and Uzuner, Ozlem: A Survey of Offensive Language Detection for the Arabic Language. ACM Transactions on Asian and Low-Resource Language Information Processing, pages 1–44, (2021)
27. Sharlene Chadwick: Impacts of Cyberbullying, Building Social and Emotional Resilience. Springer Science & Business Media, 89 pages; (2014).
28. World Health Organization, https://www.who.int/health-topics/depression.
29. Patricia Ainsworth: Understanding Depression. Univ. Press of Mississippi, 174 pages, (2000).
30. Baker DA, Algorta GP. The Relationship Between Online Social Networking and Depression: A Systematic Review of Quantitative Studies. Cyberpsychol Behav Soc Netw. (2016)
31. Moshe Isaac, Terhorst Yannik, Opoku Asare Kennedy, Sander Lasse Bosse, Ferreira Denzil, Baumeister: Predicting Symptoms of Depression and Anxiety Using Smartphone and Wearable Data. Frontiers in Psychiatry, 43 pages,(2021).
32. Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, Jun Long.: A lexicon-based approach for hate speech detection. International Journal of Multimedia and Ubiquitous Engineering. 215-230 (2015).
33. Valentino Santucci, Stefania Spina, Alfredo Milani, Giulio Biondi, Gabriele Di Bari.: Detecting hate speech for Italian language in social media. EVALITA 2018, co-located with the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018). 2263 (2018).
34. Piotr Bojanowski, Edouard Grave, Armand Joulin and Tomas Mikolov (2016). Enriching Word Vectors with Subword Information. arXiv preprint arXiv:1607.04606.
35. Gabriel Araújo De Souza, Márjory Da Costa-Abreu.: Automatic offensive language detection from Twitter data using machine learning and feature selection of metadata. International Joint Conference on Neural Networks (IJCNN). 20006646 (2020).
36. Hamada A Nayel, HL Shashirekha.: DEEP at HASOC2019: A Machine Learning Framework for Hate Speech and Offensive Language Detection. Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). 2517 (2019).
37. Zesis Pitenis, Marcos Zampieri, Tharindu Ranasinghe.: Offensive Language Identification in Greek. European Language Resources Association. 5113–5119 (2020).
38. Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, Ritesh Kumar.: Predicting the Type and Target of Offensive Posts in Social Media. arXiv preprint arXiv:1902.09666. (2020).

39. Baptist Vandersmissen: Automated detection of offensive language behavior on social networking sites. IEEE Transaction, (2012).
40. Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, Vasudeva Varma.: Deep learning for hate speech detection in tweets. Proceedings of the 26th international conference on World Wide Web companion. 759-760 (2017).
41. Taufic Leonardo Sutejo, Dessi Puji Lestari.: Indonesia Hate Speech Detection Using Deep Learning. International Conference on Asian Language Processing (IALP). 18417647 (2018).
42. Hamdy Mubarak, Kareem Darwish, Walid Magdy.: Abusive language detection on Arabic social media. Proceedings of the first workshop on abusive language online. 52-56 (2017).
43. Batoul Haidar, Maroun Chamoun, Ahmed Serhrouchni.: Offensive A Multilingual System for Cyberbullying Detection: Arabic Content Detection using Machine Learning. Adv. Sci. Technol. Eng. Syst. J. 2(6). 275-284 (2017).
44. Raghad Alshaalan, Hend Al-Khalifa.: Hate Speech Detection in Saudi Twittersphere: A Deep Learning Approach. Applied Sciences. Multidisciplinary Digital Publishing Institute. 8614 (2020).
45. Ibrahim Abu Farha, Walid Magdy.: Multitask Learning for Arabic Offensive Language and Hate-Speech Detection. Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection. 86-90 (2020).
46. Francielle Alves Vargas, Fabiana Rodrigues de Góes, Isabelle Carvalho, Fabrício Benevenuto, Thiago Alexandre Salgueiro Pardo: Contextual Lexicon-Based Approach for Hate Speech and Offensive Language Detection. ArXiv:2104.12265, (2021).
47. Fatemah Husain.: Arabic Offensive Language Detection Using Machine Learning and Ensemble Machine Learning Approaches. arXiv preprint arXiv:2005.08946. (2020).
48. Hatem Haddad, Hala Mulki, Asma Oueslati.: T-hsab: A tunisian hate speech and abusive dataset. International Conference on Arabic Language Processing. 251-263 (2019).
49. Imane Guellil, Ahsan Adeel, Faical Azouaou, Mohamed Boubred, Yousra Houichi, Akram Abdelhaq Moumna: Sexism detection, ArXiv:2104.01443, (2021).
50. Nikhil Singh, Paras Ahuja: Fundamentals of Deep Learning and Computer Vision. BPB Publications, 181 pages, (2020).
51. Raghav Aggarwal, https://medium.com/@raghavaggarwal0089/bi-lstm-bc3d68da8bd0. (2019).
52. Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov, https://research.fb.com/blog/2016/08/fasttext/. (2016).
53. Robert Nisbet, John Elder, Gary Miner: Chapter 8 - Advanced Algorithms for Data Mining. Handbook of Statistical Analysis and Data Mining Applications, Academic Press, Pages 151-172, (2009).
54. Jake VanderPlas: Python Data Science Handbook. O'Reilly Media, 548 pages, (2016).