PEOPLES DEMOCRATIC REBUPLIC OF ALGERIA

MMINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCHE

UNIVERSITY MOHAMMED BOUDIAF - M'SILA

FACULTY: MATHEMATICS AND COMPUTER SCIENCES

DEPARTEMENT: COMPUTER SCIENCES

N°: …….

DOMAIN: MATHEMATICS AND COMPUTER SCIENCES

BRANCH: COMPUTER SCIENCES

OPTION: RTIC

Dissertation submitted to obtain Master Degree

PRESENTED BY

BILAL BENTOUMI

ZAKARYA GHANAI

SUBJECT

SENTIMENT ANALYSIS BASED ON DEEP LEARNING

Publicly defended before the jury composed of

Supervisor

Mm. Ahlem BOUZAROURA University of M'sila

Academic Year 2020/2021

Contents

2

.

3

List of Figures

GENERAL INTRODUCTION

The Internet has become an essential tool for the exchange of information, both on a personal and professional level. The Web offers us a stupendous world of information, it has evolved from a simple set of static pages to increasingly complex web services and applications. Currently, the web offers us the opportunity to buy all products, read our favorite newspaper online, chat in multiple forums or have the opportunity to express ourselves on blogs.

Therefore, for most of us, the web is a great portal to find information, book a plane or a hotel, buy products, consult the opinions of other users on the products that interest us, read the comments before choosing the film to see in the cinema, see suggestions from other people before choosing the gifts, etc. the problem now, not with the internet but rather with sociological considerations, is that globalization is invading us, specifically with the emergence of social networks which have become a space where people freely express their feelings. Commercial companies took advantage of this to expose their product to critics to get constant feedback.

A new field was born, Sentiment Analysis and Opinion Mining. This domain is the descendant of its famous ancestor: data mining, better known as Data Manning. Like the latter, it aims to seek relevant information and patterns in the opinions, points of view and opinions of Internet users.

Sentiment analysis can be applied in many fields, specifically the field of e-commerce. but it can be difficult to implement, as human language is complex to interpret for machine-based learning systems, the processing of feelings becomes more complex, resulting in the need to employ other areas. such as: natural language automatic processing, machine learning.

Goal

The objective of this thesis will be to automatically detect the feelings of Internet users and their opinions for or against a product or a social phenomenon. By developing a sentiment analysis system to classify opinions into three categories: positive, negative and neutral, using deep learning methods

CHAPTER 1

SENTIMENT ANALYSIS AND OPINION MINING

1. Introduction

Sentiment analysis is a similar concept to emotion analysis, but it generally goes less far in the analysis and is often satisfied with determining three levels (favorable, neutral, unfavorable) with regard to the brand, the individual or the event.

Sentiment analysis requires much more understanding of the language than text analysis and classification by subject. Indeed, if the simplest algorithms consider only the statistics of frequency of appearance of words, this is generally insufficient to define the dominant opinion in a document, especially when the content is short like messages in a forum or tweets.

2. Sentiment Analysis and Opinion Mining

Sentiment analysis is an approach mainly based on text mining and semantic analysis that allows to determine the "position" of the individuals studied with regard to a brand or an event. Sentiment analysis can, however, also be based on other elements than textual data. It can for example be based on the use of emoticons (emojis), on Facebook "emotions", on voice analysis or even on facial coding / decoding. When based on textual mining, sentiment analysis can focus on verbatim reports from social networks, reviews, forums, qualitative or quantitative study data, etc. Feelings are generally classified into three types: negative, neutral, and positive. [1]

3. Levels of Sentiment Analysis and Opinion Mining

The sentiment analysis is done according to levels and objective:

- At document level

determines the general opinion of the whole document. This analysis works well for documents that present a specific point of view, but less so for comparisons because it will not differentiate between the topics discussed.

- At the sentence level

determines the general opinion of a sentence (positive, negative or neutral). This analysis can give a measure of the "neutrality" of a text for example to analyze Wikipedia entries. The methods used are that of the analysis of subjectivity.

- In terms of aspects (called Feature level)

instead of determining the entities to be analyzed based on structural criteria (sentence, paragraph, document), these methods are based on the correlation analysis between the opinion expressed and the target of this opinion. For example, the sentence "The subject of the course fascinates me but the teacher is boring." presents two feelings about the entity "course": the subject who is perceived as positive and the teacher, who is perceived as negative. This level of analysis makes it possible to differentiate the aspects which are liked or not by the authors of the texts and thus makes it easier to determine possible remedies. On the other hand, it is very difficult to set up because this type of analysis is extremely complex. [2]

4. Sentiment Analysis and Opinion Mining process

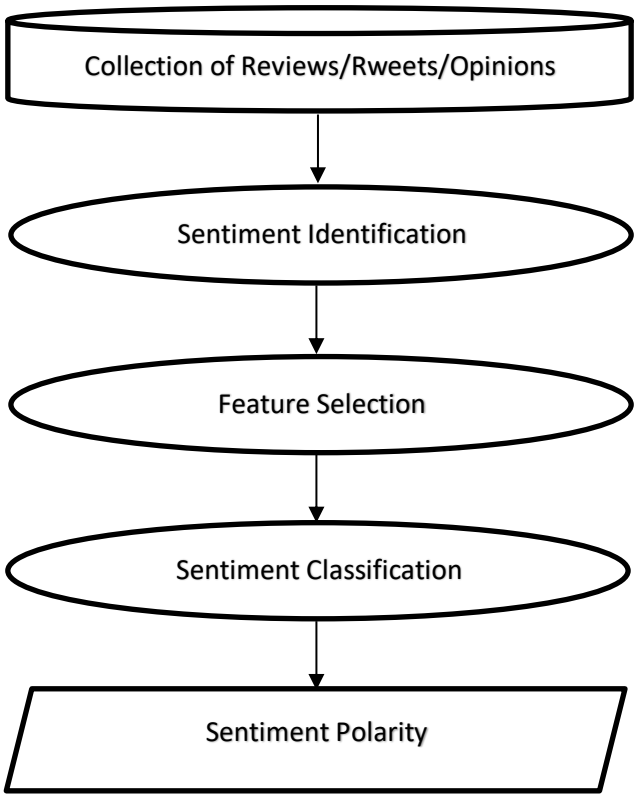

Figure 1: Sentiment analysis process. [3]

5. Disciplines related to Sentiment Analysis and Opinion Mining

Several disciplines have a direct or less direct relationship with sentiment analysis and opinion mining. Artificial intelligence, automatic natural language processing, text mining and even data mining offer indispensable tools and algorithms for the processing and classification of feelings.

5.1 Text Mining

Text mining is the analysis of data contained in a text in natural language. Applying text mining techniques to solve business problems is called text analysis.

Text mining can help an organization extract potentially valuable business information from textual content such as Word documents, emails, and social media posts such as Facebook, Twitter, and LinkedIn. Exploring unstructured data with Natural Language Processing (NLP), statistical modeling, and machine learning techniques, however, can be difficult, as natural language text is often inconsistent. It contains ambiguities caused by inconsistent syntax and semantics, including slang, language specific to vertical industries and age groups, double meanings, and sarcasm.

Text analysis software can help by translating words and phrases in unstructured data into numeric values which can then be linked to structured data in a database and analyzed with traditional data mining techniques. With an iterative approach, an organization can successfully use text analysis to better understand content-specific values such as sentiment, emotion, intensity, and relevance. Because text analytics technology is still considered an emerging technology, the results and depth of analysis can vary wildly from vendor to vendor. [4]

5.2 Natural Language Processing

Natural Language Processing (NLP) is a way for computers to analyze, understand and derive the meaning of human language in an intelligent and useful way. Using NLP, developers can organize and structure knowledge to perform tasks such as automatic synthesis, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and speech recognition. segmentation of headings.

"Apart from common word processing operations which treat text as a simple sequence of symbols, NLP considers the hierarchical structure of language: several words form a sentence, several sentences form a sentence and, ultimately, sentences which convey meaning. ideas. " Meltwater Group expert said in How Natural Language Processing Helps Uncover Social Media Sentiment. "By analyzing language for its meaning, NLP systems have long performed useful roles, such as correcting grammar, converting speech to text, and machine translation between languages."

NLP is used to analyze text, allowing machines to understand how humans speak. This human-computer interaction enables real applications such as automatic text synthesis, sentiment analysis, topic extraction, named entity recognition, labeling of parts of speech, relationship extraction, stemming, etc. NLP is commonly used for text mining, machine translation, and automatic question answering.

NLP algorithms are generally based on machine learning algorithms. Instead of manually coding large sets of rules, NLP can rely on machine learning to automatically learn these rules by parsing a set of examples (a large corpus, like a book, down to a collection of sentences) and making a static inference. In general, the more data analyzed, the more accurate the model will be. [5]

- **Latent Semantic Analysis**

Latent semantic analysis (LSA, from English: Latent semantic analysis) or latent semantic indexation (or LSI, from English: Latent semantic indexation) is a process of treatment of natural languages, within the framework of vector semantics . LSA was patented in 1988 and published in 1990.

It makes it possible to establish relationships between a set of documents and the terms they contain, by constructing “concepts” linked to the documents and to the terms.

5.3 Machine Learning

Machine learning (in English, machine learning, literally "machine learning"), field of study of artificial intelligence, concerns the design, analysis, development and implementation of methods allowing a machine (in the broad sense) to evolve by a systematic process, and thus to fulfill difficult or problematic tasks by more traditional algorithmic means.

The analysis can concern graphs, trees, or curves (for example, the curve of temporal evolution of a measure; we then speak of continuous data, as opposed to discrete data associated with classic attributes-values) in the same way. than simple numbers. [6].

The algorithms used allow, to a certain extent, a computer-driven (possibly a robot), or computer-assisted system, to adapt its analyzes and behavior in response, based on the analysis of empirical data from 'a database or sensors.

The difficulty lies in the fact that the set of all the possible behaviors taking into account all the possible inputs quickly becomes too complex to describe (we speak of a combinatorial explosion). Programs are therefore entrusted with the task of adjusting a model to simplify this complexity and to use it in an operational manner. Ideally, the learning will aim to be unsupervised, i.e. the nature of the training data is not known.

These programs, depending on their degree of sophistication, may include probabilistic data processing capabilities, data analysis from sensors, recognition (voice recognition, pattern recognition, handwriting, etc.), data-mining, 'theoretical computer science… [7]

When data is not labeled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find a natural grouping of data to groups, and then map new data to those groups. trained. Support Vector Cluster Algorithm created by Hava Siegelmann and Vladimir Vapnik applies support vector statistics developed in Support Vector Machine Algorithm to categorize unlabeled data and is one of the most widely used clustering algorithms in industrial applications. [6]

6. Sentiment Analysis and Opinion Mining classification methods

Sentiment analysis and opinion mining is essentially based on the classification of feelings into three categories (positive, negative, and neutral) according to [8] there are two main approaches:

- ✓ Machine learning-based approach
- ✓ NLP-based approach

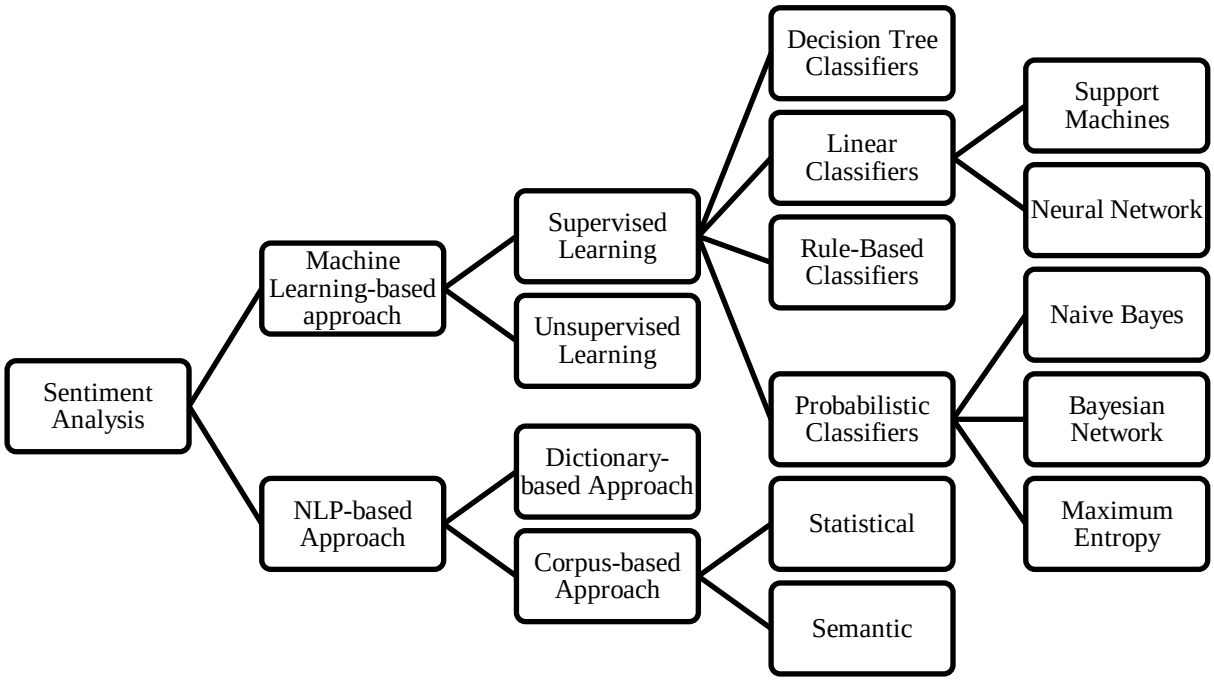

Figure 2: Sentiment classification methods. [3]

The techniques of machine learning are known, in what follows we will detail the NLP-based approach which will be the heart of our work:

6.1 Approach based on Natural Language Processing

NLP is based under sentence processing, while sentences are made up of words. In sentiment analysis we find words that are used to express opinion, a specific lexicon or vocabulary is used in the classification of sentiment. There are positive opinion words that are used to express positive feelings such as: I like it; beautiful, pretty, interesting… and negative opinion words which are used to express negative feelings. There are also opinion phrases and idioms which together are called opinion lexicon. There are three main approaches to compiling or collecting the opinion word list.

The manual approach is very time consuming and is not used alone. It is usually combined with the other two automated approaches as a final check to avoid errors that have resulted from automated methods. The two automated approaches are as follows. [8]

- Dictionary-based approach

A small set of opinion words are collected manually with known directions. Then, this set is cultivated by research in the well-known WordNet corpora or thesauri for their synonyms and antonyms. The newly found words are added to the start list then the next iteration begins. The iterative process stops when no new word is found. After the process is complete, a manual inspection can be performed by an expert to remove or correct the errors.

- Corpus-based approach

The corpus-based approach helps solve the problem of finding opinion words with context-specific directions. Its methods depend on syntactic patterns or patterns that occur together with a corpus of opinion words to find other opinion words in the same context. One of these methods was represented by Hatzivassiloglou and McKeown. They started with a list of opinion seed adjectives, and used them with a set of language constraints to identify other opinion adjective words and their orientations. Constraints are for connectors like AND, OR, BUT, EITHER OR-OR. . .... the conjunction AND for example says that the conjoined adjectives generally have the same orientation. This idea is called the consistency of feeling, which is not always consistent in practice.

There are also adversarial expressions such as but, however, which are indicated as changes of opinion. In order to determine if two conjoined adjectives are the same or different orientations, the learning is applied to a large corpus. Then, the links between the adjectives form a graph and the grouping is done on the graph to produce two sets of words: positive and negative. [8]

6.2 Approach based on Machine Learning

In machine learning, supporting vector machines (SVM, also supporting vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm constructs a model that assigns new examples to one category or another, this which makes it a non-probabilistic linear binary classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples in the separate categories are divided by as large a free space as possible. New examples are then mapped into that same space and predicted to fall into a category based on which side of the gap they fall.

In addition to performing linear classification, SVMs can efficiently perform non-linear classification using what is called the kernel trick, by implicitly matching their inputs into large-dimensional feature spaces.

7. Sentiment Analysis Tools

There are tools to identify the feeling given off by a text. Here is a non-exhaustive list of the most famous tools:

- Werfamous

Free online analysis tool, giving a sentiment score on a scale of -100 to 100, as well as a level of confidence linked to this score

The site saves text trackers (cookies) on your device in order to guarantee you better content and for statistical collection purposes. You can deactivate the use of cookies by changing your browser settings. By continuing to browse our site without changing your browser settings you are giving us permission to store information on your device.

You place your query in the dialog box and get an analysis on the web. Queries are saved in their Dataset and History sections. [9].

- AFINN

AFINN is a list of words valid for valence with an integer between negative five (negative) and plus five (positive).

Assesses the positivity / negativity of a word using a dictionary contained in an archive.

Although the title of the associated document suggests that it is based on the corpus labeled ANEW, it is not. The title is simply a wordpun. It was developed independently of the wordlist, and this is not a review of it. Compared to ANEW, AFINN wordlist has more words and includes obscene words. ANEW on the other hand has (besides valence) excitement and dominance for every word and every word has been tagged by

multiple people and the mean and standard deviation are given. AFINN has only been tagged by Finn Årup Nielsen. Finn Årup Nielsen was in no way involved in the development of ANEW. ANEW was developed by Margaret M. Bradley and Peter J. Lang. [10]

- General Inquirer

Lemmatizes words, performs graphical and statistical analysis, and produces a report containing sentences with the most significant words.

This site is divided into several sections, giving both information about the Inquirer and points to other systems. The pages of our website contain links to the first 100 words of each category in the Harvard and Lasswell dictionaries. [11]

- SenticNet

Talking about SenticNet is talking about sentiment analysis at the conceptual level, that is, performing tasks such as detecting polarity and recognizing emotions relying solely on semantics and linguistics.

In this context, SenticNet can be any of the following:

- ✓ a conceptual knowledge base
- ✓ a multidisciplinary framework
- ✓ a private company.

As a knowledge base, SenticNet provides a set of semantics, sentics, and polarity associated with 100,000 natural language concepts. In particular, semantics is the most semantically related to the concept of entry (i.e. the five concepts that share more semantic characteristics with the concept of entry), sentics are values of categorization of emotions expressed in terms of four affective dimensions (Pleasant, Attention, Sensitivity, and Aptitude) and polarity is a number floating between -1 and +1 (where -1 is extreme negativity and +1 is extreme positivity). The Knowledge Base is free to download as a stand-alone XML file, and its latest version (released every two years) is also accessible as an API.

As a framework, SenticNet consists of a set of tools and techniques for sentiment analysis combining common sense reasoning, psychology, linguistics and machine learning. In this context, SenticNet is more commonly referred to as "seismic informatics," a multidisciplinary paradigm that goes beyond simple statistical approaches to focus on a representation preserving the semantics of natural language concepts and on sentence structure. [12]

- SentiWordNet

SentiWordNet is a lexical resource for opinion mining. SentiWordNet assigns each WordNet synset three sentiment scores: positivity, negativity, objectivity.

In the area of sentiment analysis, a comparative study was carried out to determine the advantages and disadvantages of each data source. As part of the analysis of tweets relating to major events, the study highlights the fact that several of these tweets could not be recognized by the data sources. It can be seen that SentiWordNet, SenticNet and SentiStrength seem to cover a greater number of tweets. However, the article also highlights that the coverage rate is not synonymous with efficient recognition and that the polarity of a given word is unreliable. This is why the article proposes to combine several of these methods in order to exploit the advantages of each and to obtain the result as close as possible to reality. [13]

- Sentiment140

Sentiment140 (formerly known as "Twitter sentiment") is a free online tool that was created by three computer science students at Stanford, so this is an academic project. This tool, unlike most other sentiment analysis sites, does not use positive or negative word lists but is based on machine learning algorithms. [14]

Sentiment140 allows you to discover feelings of tweets of a brand, a product or a subject on Twitter. Official Website: http://www.sentiment140.com.

- Tweetfeel

Tweetfeel is a service that leverages the real-time capabilities of Twitter to give you Twitter user feedback on a keyword, brand, or star.

TweetFeel is evaluated on the basis of the presence of specific keywords in tweets such as Good, Bad, etc ... (English only for the moment), when a similar service in French.

Then a percentage is calculated according to the number of tws or negative an overall feeling of Twitter on the brand. [15] Official Website: http://www.tweetfeel.com.

- Twitrratr

Twitrratr is a free online tool, which emerged from a Startup Weekend project. Twitrratr works from a list of positive words and a list of negative words [16].

This tool ranks an opinion on the query keyword if it is able to match it with a word from one of the two lists. Positive and negative words used to classify tweets are highlighted in the interface [17]. Official Website: http://twitrratr.com.

- Tweet Sentiment Analysis

Tweet Sentiments Analyzes is a free, open-source, online sentiment analysis tool for Twitter. It can give positive, negative and neutral feelings of tweets on the keyword thrown in the query. It can work on 12 languages. It gives the results in graphical form. [18] Official Website: http://smm.streamcrab.com.

8. Conclusion:

In this chapter we have provided an overview of the main fundamentals, methods and techniques of sentiment classification and opinion mining. in the next chapter we will detail more the work done in this field in relation to the automatic processing of natural language and present our feeling classification system.

Chapter 2

Machine Learning and Deep Learning

1. Introduction

To make our works easier, human want to create machines has same thinking of them. He tries to create a machine that imitates humane brain and his intelligence. In 1943 Pitts McCulloch invent the first Mathematical model of neurons after that the artificial neural networks had began, then Rosenblatt created a single neuron for classification, Perceptron learning rule, Perceptron convergence theorem in 1958. So, a perceptron in context of neural networks is an artificial neuron which works similar to the biological neuron and uses weights in order to classify the inputs. Machine Learning is the core subarea of artificial intelligence.

One year ago (2018), the Turing Award has been given to a godfathers of AI. Yoshua Bengio, Geoffrey Hinton, and Yann LeCun, they had invented the deep learning that requires an extensive and diverse set of data contrary to machine learning that needs fewer data to train the algorithms. Deep learning is better than machine learning because in deep learning methods include multi-layer processing with less time and better accuracy performance. Sub sampling layers give better result, by use of CNN.

2. Machine Learning

Machine learning has several types where each type used for specific problem. It used to solve different types of problems. It provides effective solutions to several problems.

2.1. Machine Learning Definition

Machine Learning is a subset of artificial intelligence (AI), machine learning (ML) is the area of computational science that focuses on analyzing and interpreting patterns and structures in data to enable learning, reasoning, and decision making outside of human interaction. Simply put, machine learning allows the user to feed a computer algorithm an immense amount of data and have the computer analyze and make data-driven recommendations and decisions based on only the input data. If any corrections are identified, the algorithm can incorporate that information to improve its future decision making. [19]

Machine learning is the science of getting computers to act without being explicitly programmed machine learning is so pervasive today that you probably use it dozens of times

a day without knowing it. Many researchers also think it is the best way to make progress towards human-level AI [20].

2.2. Usage of Machine Learning

- ✓ Problems for which existing solutions require a lot of hand-tuning or long lists of rules: one Machine Learning algorithm can often simplify code and perform better [20].
- ✓ Complex problems for which there is no good solution at all using a traditional approach: the best Machine Learning techniques can find a solution [21].
- ✓ Fluctuating environments: a Machine Learning system can adapt to new data [21].
- ✓ Getting insights about complex problems and large amounts of data [21].

2.3. Types of Machine Learning Systems

There are four types of machine learning systems, which are classified by the type and amount of supervision they need during the learning process [21]:

- ✓ Supervised learning.
- ✓ Unsupervised learning.
- ✓ Semi supervised learning.
- ✓ Reinforcement Learning.

2.3.1. Supervised learning

The supervised learning means the training data you feed to the algorithm includes the desired solutions, called labels [21], see **Figure 3**.

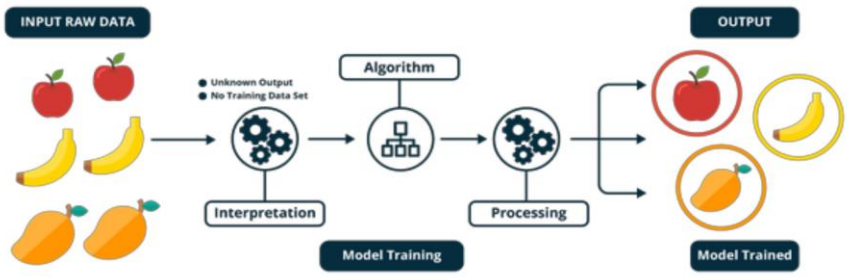

Figure 3: Example of Supervised Learning.

Figure 3 is an example of machine learning since the outputs are known in advance, namely apples, bananas and kiwi, the machine here learns through examples and solutions.

The most important supervised learning algorithms:

- ✓ k-Nearest Neighbors
- ✓ Linear Regression
- ✓ Logistic Regression
- ✓ Support Vector Machines (SVMs)
- ✓ Decision Trees and Random Forests
- ✓ Neural networks.

- K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a supervised machine learning algorithm based on the distance between items. The nearest neighbors to this item is considered to be its own category, where K is the number of the nearest neighbors.

Example: as shown in **Figure 4**:

Our item is the point, if the value of K=1 the 1-nearest neighbor of the point is a plus. If the value of K=2 the 2-nearest neighbor of the point is unknown, because the both of the plus and the minus signs achieve the same score. If the value of K=5 the 5-nearest neighbor of the point is a minus [22].

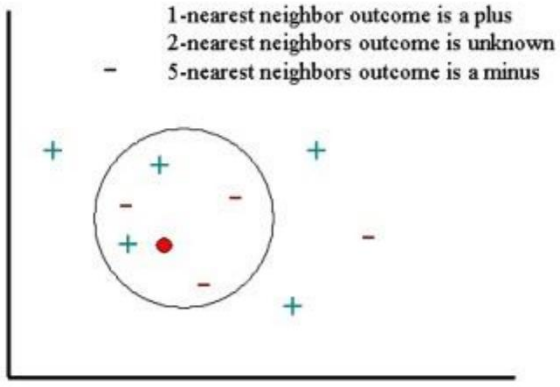

Figure 4: K-Nearest Neighbors.

- Support Vector Machines

Support vector machine (SVM) is a supervised machine learning algorithm based on finding a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points. To separate the two classes of data points, there are many possible hyperplanes that could be chosen. The objective of SVM is to find a

plane that has the maximum margin, i.e. the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence [23], see **Figure 5**.

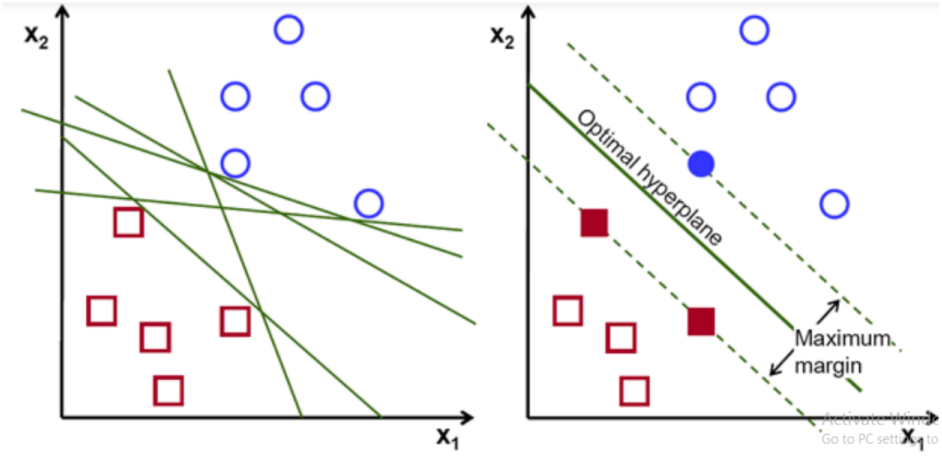

Figure 5: Possible hyperplanes.

2.3.2. Unsupervised learning

In unsupervised learning, as you might guess, the training data is unlabeled, the system tries to learn without a teacher [21].

The most important unsupervised learning algorithms:

- Clustering
  - ✓ k-Means
  - ✓ Hierarchical Cluster Analysis (HCA)
  - ✓ Expectation Maximization
- Visualization and dimensionality reduction
  - ✓ Principal Component Analysis (PCA)
  - ✓ Kernel PCA
  - ✓ Locally-Linear Embedding (LLE)
  - ✓ t-distributed Stochastic Neighbor Embedding (t-SNE)

- Association rule learning
  - ✓ Apriori
  - ✓ Eclat

2.3.3. Semi-Supervised Learning

Some algorithms can deal with partially labeled training data, usually a lot of unlabeled data and a little bit of labeled data, this is called semi supervised learning. Often, these algorithms are a combination of algorithms supervised algorithms with others unsupervised algorithms [21].

2.3.4. Reinforcement Learning

Reinforcement Learning is a very different beast. The learning system, called an agent in this context, can observe the environment, select and perform actions, and get rewards in return or penalties in the form of negative rewards [21].

2.4 Main challenges of Machine Learning

There are many challenges in Machine Learning, the most important are:

- ✓ Insufficient Quantity of Training Data.
- ✓ Nonrepresentative Training Data.
- ✓ Poor-Quality Data.
- ✓ Irrelevant Features.
- ✓ Overfitting the Training Data.
- ✓ Underfitting the Training Data.

3. Deep Learning

Artificial intelligence is defined as the intelligence shown by machines and programs that mimics human mental abilities and modes of action, such as the ability to learn, infer, and react to situations that have not been programmed into the machine. The name of an academic field on how to make computers and programs capable of intelligent behavior. Senior researchers define artificial intelligence as “the study and design of intelligent systems that absorb their environment and take actions that increase their chances of success.” [24].

John McCarthy [24] who developed the term in 1955, defines it as "the science and geometry of making intelligent machines."

In the last few years, the evolution of artificial intelligence technology has leaped forward. Deep learning is a major feature of the development of artificial neural networks that mimic the way the human brain works, which is capable of experimenting and learning and developing itself without human intervention.

Deep Learning has proven its ability to recognize images, understand speech, translate from one language to another, and other abilities. There are many deep learning algorithms, but since our work will be about image classification. We will talk about the most popular learning algorithms in this field, Convolutional Neural Network (CNN) is one of the most popular algorithms of deep learning in image classification, and it has achieved many remarkable successes in this field.

3.1. Artificial Neural Networks

Artificial neural network (ANN) is an implied model of the biological neuron to make decisions and conclusions by simulating human brain's work [25]. **Figure 6** Represents an example of ANN witch has three main layers where the first layer is the input layer, the second layer is the hidden layer and the last layer is the output layer.

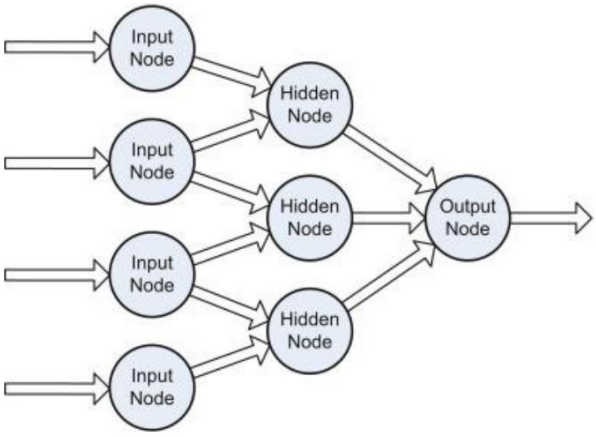

Figure 6: Artificial Neural Network.

3.2. Convolutional Neural Networks

Convolution Neural Network (CNN or ConvNet) is one of the most famous algorithms of deep learning in the field of image classification, it is a multi-layer algorithm. The first part of those layers is to extract the features of images, it consists of two basic types of layers, the first type is convolution layer and the second is pooling layer. The second part of CNN is a neural network and its task is classification of features.

3.2.1 Convolutional Neural Network Layers

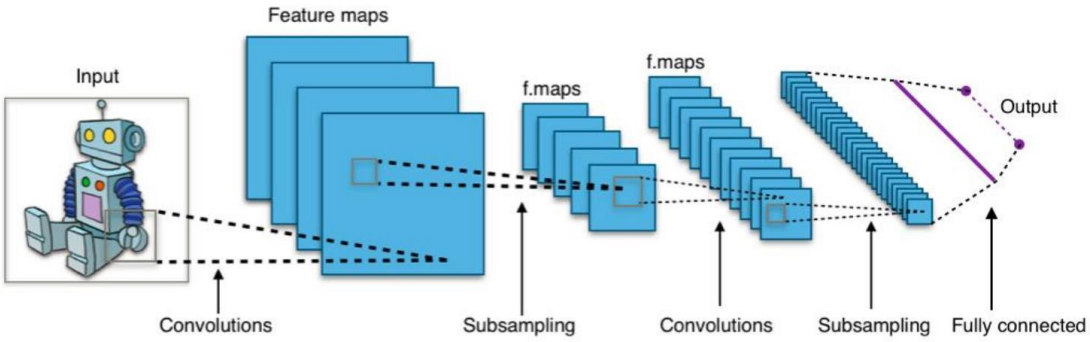

Figure 7: Convolutional Neural Network main layers.

Convolutional Neural Network is a sequence of layers, and each layer of a ConvNet transforms a volume of activations to another layer through a differentiable function, ConvNet has three main types of layers **Convolutional Layer**, **Pooling Layer**, and **Fully-Connected Layer**, see **Figure 7**. ConvNet converts the input image across different layers until we finally get the class scores, each layer may has another parameters.

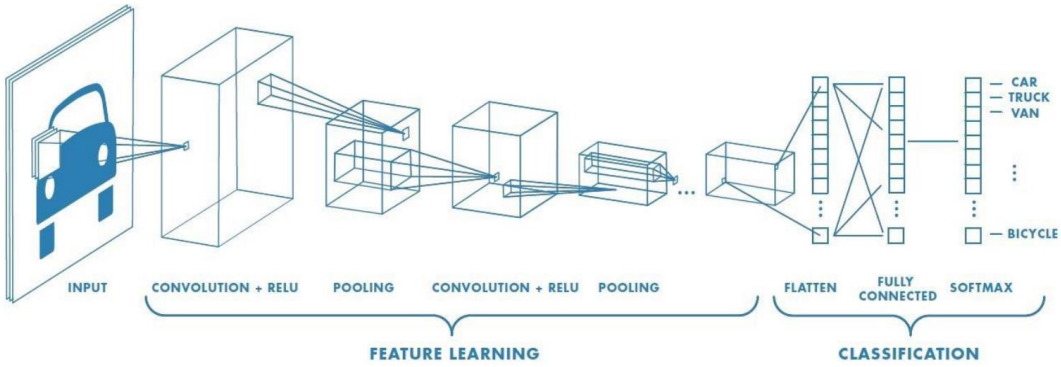

Figure 8: Convolution Neural Network Layers.

We can divide the network into two parts. The first part (features extraction part) contains a sequence of **Convolutional Layers**, **Pooling Layers**, and the second part contains a sequence of **Fully-Connected** layers. The first part (classification layer) is to extract the features of the image and the second part is to classify the image according to the output of the first part, see **Figure 8**.

- Convolutional Layer

The Convolutional layer is the core building block of a Convolutional Network that does most of the computational heavy lifting, the CONV layer's parameters consist of a set of learnable filters. Every filter is small spatially (along width and height), but extends through the full depth of the input volume [26].

- Pooling Layer

A layer that often comes after the convolutional layer. Its function is to progressively reduce the spatial size of the representation to reduce the amount of parameters and computation in the network, and hence to also control overfitting. The Pooling Layer operates independently on every depth slice of the input and resizes it spatially [26].

- Fully Connected Layer

Neurons in a fully connected layer have full connections to all activations in the previous layer, as seen in regular Neural Networks. Their activations can hence be computed with a matrix multiplication followed by a bias offset [26].

- Normalization Layer

Many types of normalization layers have been proposed for use in ConvNet architectures, sometimes with the intentions of implementing inhibition schemes observed in the biological brain. However, these layers have since fallen out of favor because in practice their contribution has been shown to be minimal [27].

Dropout is a technique used to improve over-fit on neural networks. Basically during training half of neurons on a particular layer will be deactivated. This improve generalization because force your layer to learn with different neurons the same "concept". During the prediction phase the dropout is deactivated [27], see **Figure 9**.

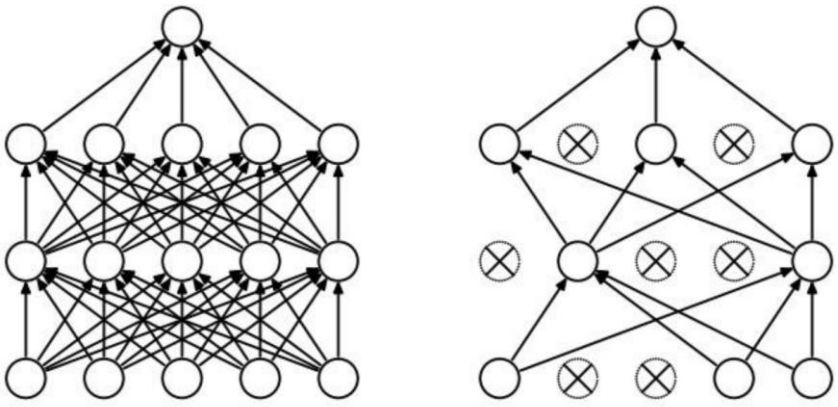

Figure 9: Dropout Layer.

3.2.2. Activation Function

Activation function is a node that is put at the end of or in between Neural Networks. They help to decide if the neuron would fire or not [26].

- Rectified Linear Units (RELU) Activation Function

is non-linear activation function. It converts the input values to positive values [26].

$$RELU = MAX(X,0)$$

X: the input values

- Softmax Activation Function

Able to handle multiple classes only one class in other activation functions—normalizes the outputs for each class between 0 and 1, and divides by their sum, giving the probability of the input value being in a specific class. Typically Softmax is used only for the output layer, for neural networks that need to classify inputs into multiple categories [26].

4 Difference between Deep Learning and Machine Learning

The major difference between deep learning and machine learning is its execution as the size of data increases. Deep learning algorithms need a large of data, when the data is small those algorithms don't perform that well. [28]

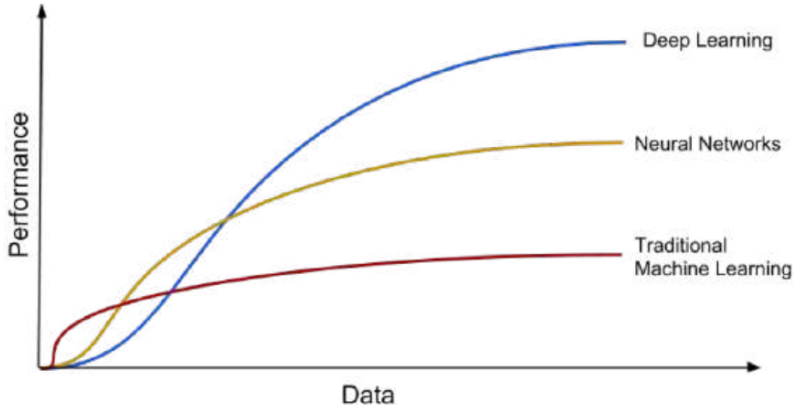

Figure 10: Machine Learning vs Deep Learning.

5 Conclusion

In this chapter we described briefly the definition of Machine Learning and Deep learning, we have mentioned its process and its importance and the major difference between deep learning and machine learning we also have explained CNN architecture and the different layer and the most popular types of activation functions of CNN.

CHAPTER 3

DESIGN AND PRODUCTION

1. Introduction

Nowadays, Opinion Mining has become an emerging research topic due to the large amount of data available on blogs and social networks. Tracking different types of opinions and summarizing them can provide valuable information of different types of opinions to users who use social media to get feedback on any product, service or topic. Analyzing opinions and classifying it on the basis of polarity (positive, negative, neutral) is a difficult task. Much work has been done on sentiment analysis of Twitter data and a lot needs to be done.

2. Related Works

Several works are carried out to classify the data in this section we will present some following the main axes:

2.1. Sentiment analysis based on LSA (2010)

Online user reviews help consumers cope with information overload and facilitate decision making. However, many user reviews online do not have enough utility votes for other users to gauge their true level of utility. This study empirically examines the impact of different characteristics, i.e., the basic, stylistic and semantic characteristics of online user reviews on the number of utility votes these reviews receive. Text mining techniques are used to extract semantic characteristics from revision texts. The results show that semantic characteristics are more influential than other characteristics when it comes to the number of utility votes received. The results also suggest that critics with extreme opinions receive more utility votes than those with mixed or neutral opinions. This paper sheds light on the understanding of the utility voting behavior of online users and the design of a better utility voting mechanism for online user review systems. [29]

2.2. Multi-class SVM (2011)

The ubiquity of Web2.0 makes the Web an invaluable source of business information. For example, the product Reviews co-authored by many independent Internet reviewers can help consumers make purchasing decisions and enable businesses to improve their business strategies. As the number of reviews is growing exponentially, exploration and opinion research techniques are required to identify important reviews and opinions to answer user questions. Most opinion mining and extraction approaches attempt to extract sentimental or

bipolar expressions from a large volume of reviews. However, the process often ignores the quality of each revision and can pick up unnecessary or even noisy documents. In this article, we propose a method for evaluating the quality of information in product reviews. We treat review quality assessment as a classification problem and employ an effective information quality framework to extract representative review characteristics. Experiments based on a body of expert data demonstrate that the proposed method significantly outperforms state-of-the-art approaches. [30]

2.3. Lexicon Based Sentiment Analysis, NLP (2012)

Highlights - Ability to deal with the tendency of many users to express their opinions in non-standard language. Detects user opinion targets in multi-domain scenarios. Modular language knowledge model with low cost adaptability. Hierarchical lexicon specially designed to analyze news comments. Thanks to the technological revolution that has accompanied Web 2.0, users can interact intensively on the Internet, as evidenced by social networks, blogs, forums, etc. In these scenarios, users can express themselves freely on any relevant topic. However, the high volume of user-generated content makes manual analysis of this discourse unsustainable. Therefore, automatic analysis techniques are needed to extract the opinions expressed in user reviews, as these opinions are an implicit barometer of unquestionable interest to a wide variety of businesses, agencies and organizations. organizations. We therefore provide a lexicon-based commentary-based News Sentiment Analyzer (LCN-SA) which is capable of handling the following: (a) the tendency of many users to express their opinions in non-standard language; (b) target detection of user opinions in a multi-domain scenario; (c) designing a modularized linguistic knowledge model with low cost adaptability. The proposed system consists of an automatic focus detection module and a sentiment analysis module capable of assessing the opinions of users on the topics covered in the news. These modules use a taxonomy-lexicon specifically designed for news analysis. Experiments show that the results obtained so far are extremely promising. [31]

2.4. Sentiment Analysis Based on Semantic Analysis (2013)

Estimation of polarity in large-scale and multi-subject domains is a difficult question. Most of the advanced solutions are essentially based on the frequencies of words carrying feelings (extracts from a lexicon, for example) when analyzing the feeling conveyed by a text in natural language. These approaches ignore the structural aspects of a document, which contain valuable information. Rhetorical Structure Theory (RST) provides important information about the relative importance of different text boxes in a document. This knowledge could be useful for sentiment analysis and polarity classification. However, RST has only been studied for polarity classification issues in constrained and small-scale scenarios. The main objective of this article is to explore the utility of RST in the large-scale polarity ranking of blog posts. We apply sentence level methods to select key phrases that express the overall sentiment of an article on the topic. Then, we apply the RST analysis to these base sentences to guide the classification of their polarity and thus generate an overall estimate of the polarity of the document with respect to a specific subject. Our results show that RST provides valuable information on the discursive structure of texts which can be used

to more accurately classify documents in terms of their estimated sentiment in multi-topic blogs. [32]

2.5. Sentiment Analysis Based on Machine Learning (2014)

Presented variations of the Naive Bayes classifiers to detect the polarity of English tweets. Two different variations of the Naive Bayes classifiers have been constructed, namely Baseline (trained to classify tweets as positive, negative, and neutral) and Binary (uses a polarity lexicon and ranks as positive and negative. Neutral tweets are neglected). The characteristics considered by the classifiers were Lemmas (nouns, verbs, adjectives and adverbs), Lexicons of Polarity and Multiword from different sources and Valence Shifters. [33]

3. Programming Languages and Libraries used

We decided to use Python programming Language due to its popularity in Artificial Intelligence field and to the huge number of Libraries.

- ✓ **Python**

  Python is an interpreted high-level general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant indentation.

- ✓ **NumPy**

  NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

- ✓ **Pandas**

  Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license.

- ✓ **Keras**

  Keras is an open-source software library that provides a Python interface for artificial neural networks. Keras acts as an interface for the TensorFlow library. Up until version 2.3, Keras supported multiple backends, including TensorFlow, Microsoft Cognitive Toolkit, Theano, and PlaidML.

4. IDEs

We worked with **PyCharm** and **WebStorm** IDEs

- ✓ **PyCharm**

  PyCharm is a dedicated Python Integrated Development Environment (IDE) providing a wide range of essential tools for Python developers, tightly integrated to create a convenient environment for productive Python, web, and data science development.

- ✓ **WebStrom**

  WebStorm is an integrated development environment for coding in JavaScript and its related technologies including TypeScript, React, Vue, Angular, Node.js, HTML, and style sheets.

5. Frameworks

We choose to work with **Electron.js** and **Vue.js** to build our Desktop Application User Interface.

- ✓ **Electron.js**

  Electron is an open-source software framework developed and maintained by GitHub. It allows for the development of desktop GUI applications using web technologies: it combines the Chromium rendering engine and the Node.js runtime.

- ✓ **Vue.js**

  Vue.js is an open-source model–view–viewmodel front end JavaScript framework for building user interfaces and single-page applications. It was created by Evan You, and is maintained by him and the rest of the active core team members.

6. Code Presentation

Our project is based on the dataset available at https://www.kaggle.com/kazanova/sentiment140 which is composed of about 1.5m tweets to train sentiment predictors.

- ✓ **Imports**

  Here we have imported the necessary libraries that we are going to use in our project.

```
import os
os.environ['TF_CPP_MIN_LOG_LEVEL'] = '3'

import re
import gensim
import numpy as np
import pandas as pd

from keras import layers, regularizers
from keras.callbacks import ModelCheckpoint
from keras.models import Sequential
from keras.preprocessing.text import Tokenizer
from keras_preprocessing.sequence import pad_sequences

from nltk.tokenize.treebank import TreebankWordDetokenizer

from sklearn.model_selection import train_test_split

from tensorflow.python.debug.examples.debug_mnist import tf
```

Figure 11: Library Imports

✓ **Preprocess Data**

A tweet contains a lot of opinions about the data that are expressed in different ways by different users. The Twitter dataset used in this study work is already labeled into three classes, negative, neutral and positive polarity and therefore the sentiment analysis of the data become easy to observe the effect of various features. Raw data with polarity is highly susceptible to inconsistency and redundancy. Preprocessing the tweet includes the following points,

- Remove all the URLs (ex: www.xyz.com), hashtags (ex: #topic), targets (ex: @username)
- Correct the spellings, the sequence of repeated characters must be processed
- Replace all emoticons with their sentiment.
- Remove all punctuation, symbols, numbers.
- Delete stop words.
- Expand Acronyms (you can use an acronym dictionary)

```
dataset = pd.read_csv('datasets/sentiment140.csv', header=None, usecols=[0,5], names=['sentiment',
'selected_text'])

temp = []

data_to_list = dataset['selected_text'].values.tolist()

# Preprocess Text
for i in range(len(data_to_list)):
    temp.append(preprocess_text(str(data_to_list[i])))

# Sentences to Words
data_words = list(sent_to_words(temp))

# Detokenize
data = []
for i in range(len(data_words)):
    data.append(detokenize(data_words[i]))
```

Figure 12: Preprocess Data

```
def preprocess_text(text):

  # Remove html tags
  text = re.sub(r'<[^>]+>', '', text)
  # Remove URLs
  text = re.sub(r'https?://\S+|www\.\S+', '', text)
  # Remove Emails
  text = re.sub('\S*@\S*\s?', '', text)
  # Remove new line characters
  text = re.sub('\s+', ' ', text)
  # Remove distracting single quotes
  text = re.sub('\'', '', text)
  # Remove punctuations and numbers
  text = re.sub('[^a-zA-Z]', ' ', text)
  # Single character removal
  text = re.sub(r"\s+[a-zA-Z]\s+", ' ', text)
  # Removing multiple spaces
  text = re.sub(r'\s+', ' ', text)

  return text

def sent_to_words(sentences):
    for sentence in sentences:
        yield(gensim.utils.simple_preprocess(str(sentence), deacc=True))

def detokenize(text):
    return TreebankWordDetokenizer().detokenize(text)
```

Figure 13: Functions Definition

✓ Data sequencing

We'll implement the Keras tokenizer as well as its pad_sequences method to transform our text data into 3D float data, otherwise our neural networks won't be able to be trained on it.

```
max_words = 5000
max_len = 200

tokenizer = Tokenizer(num_words=max_words)
tokenizer.fit_on_texts(data)
sequences = tokenizer.texts_to_sequences(data)
tweets = pad_sequences(sequences, maxlen=max_len)
```

Figure14: Data Sequencing

Output of "print(tweets)":

```
[[   0    0    0 ...   68  146   41]
 [   0    0    0 ...    0  397   65]
 [   0    0    0 ...    0    0   11]
 ...
 [   0    0    0 ...  372   10    3]
 [   0    0    0 ...   24  542    4]
 [   0    0    0 ... 2424  199  657]]
```

✓ Data splitting

We must split our dataset to 75% train, 25% test. the train part used to train the model and the test part is used to validate the accuracy of the model.

```
# Splitting the data
X_train, X_test, y_train, y_test = train_test_split(tweets,labels, random_state=0)
print (len(X_train),len(X_test),len(y_train),len(y_test))
```

Figure 15: Data Splitting

Output:

20610 6871 20610 6871

✓ **Model Building**

We have implemented a Convolutional 1D network, known as well by delivering good metrics when talking about NLP.

```
model = Sequential()

model.add(layers.Embedding(max_words, 40, input_length=max_len))
model.add(layers.Conv1D(20, 6, activation='relu', kernel_regularizer=regularizers.l1_l2(l1=2e-3, l2=2e-
3),bias_regularizer=regularizers.l2(2e-3)))
model.add(layers.MaxPooling1D(5))
model.add(layers.Conv1D(20, 6, activation='relu', kernel_regularizer=regularizers.l1_l2(l1=2e-3, l2=2e-3),
bias_regularizer=regularizers.l2(2e-3)))
model.add(layers.GlobalMaxPooling1D())
model.add(layers.Dense(3, activation='softmax'))

model.compile(optimizer='rmsprop' ,loss='categorical_crossentropy', metrics=['acc'])

checkpoint = ModelCheckpoint("model3.hdf5", monitor='val_accuracy', verbose=1,save_best_only=True, mode='auto',
period=1, save_weights_only=False)

history = model.fit(X_train, y_train, epochs=70, validation_data=(X_test, y_test), callbacks=[checkpoint])
```

Figure 16: Model Building

Output:

Epoch 1/70
645/645 [==============================] - 5s 8ms/step - loss: 1.0300 - acc: 0.553
3 - val_loss: 0.8851 - val_acc: 0.6107
Epoch 2/70
645/645 [==============================] - 5s 7ms/step - loss: 0.8474 - acc: 0.616
9 - val_loss: 0.8393 - val_acc: 0.6191
.
.
.

Epoch 70/70
645/645 [==============================] - 5s 7ms/step - loss: 0.3096 - acc: 0.917
3 - val_loss: 0.5819 - val_acc: 0.8195

Now our model is trained and we have got 91% accuracy

✓ **Testing the Model**

Here we loaded the model after is trained, to predict the sentiment of text we need to preprocess the text, sequencing it then passing it to the mode

```
text = "This is the best sentiment analysis article ever!")

model = keras.models.load_model('model.hdf5')

sentiment = ['Neutral', 'Negative', 'Positive']

sequence = tokenizer.texts_to_sequences([text])
test = pad_sequences(sequence, maxlen=max_len)
value = model.predict(test)

result = {
    'sentiment': sentiment[np.around(value, decimals=0).argmax(axis=1)[0]],
    'percent': round(np.max(value) * 100, 2)
}

print('Sentiment: ' + result['sentiment'] + ', Confidence: ' + result['percent'])
```

Figure 17: Testing the Model

Output:.

Sentiment: Positive, Confidence: 97.21%

7. Screenshots of Application

This is the main interface of our desktop application.

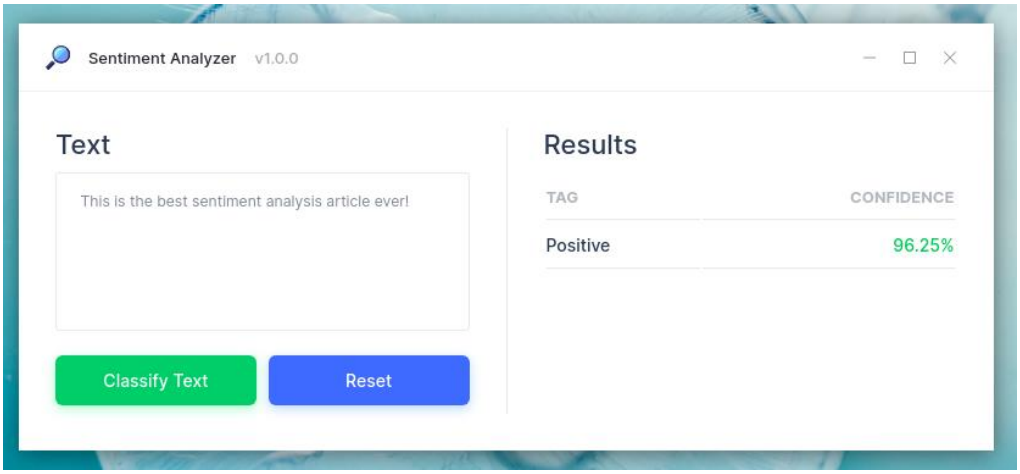

Figure 18: Application Interface

8. Conclusion

In this chapter, we have presented the essence of our work which consists in creating an sentiment analysis system to detect feelings in social networks. For the implementation, we chose the social network Twitter by examining the tweets to classify them as: Positive, negative or neutral. Our system fits into the field of artificial intelligence precisely "Deep Learning". Because the accuracy of the classification increases each time when the classification algorithm is executed.

GENERAL CONCLUSION

Sentiment Analysis and Opinion Mining is an emerging field, in recent years several research studies have focused on the task of sentiment analysis, especially in the field of microblogging.

Our work is part of this same research axis, we propose a system of subjective classification of the opinions of users of the twitter social network on a product or an event in three categories: positive, negative and neutral.

Our contribution consists in integration of deep learning methods besides Naturel Language Processing methods.

Our sentiment analysis system is still in the development phase and far from complete. For improvement, we propose the use of hybrid approaches to have better results, and possibly expand the use of this work towards other objectives such as trend analysis and knowledge extraction from social networks.

BIBLIOGRAPHY

[1] B.Bathelot, Published on 02/02/2018. https://www.definitions-marketing.com/definition/analyse-des-sentiments, Visted on 12/06/2021.

[2] Liu, B. (2014) Sentiment analysis and Opinion Mining. Morgan & Claypool Publ.

[3] https://www.researchgate.net/publication/261875740_Sentiment_Analysis_Algorithms_and_Applications_A_Survey, Visted on 12/06/2021.

[4] "Full Circle Sentiment Analysis". Breakthrough Analysis. Published on 23/02/2015.

[5] Introduction to Natural Language Processing (NLP), Published on 11/08/2016, https://algorithmia.com/blog/introduction-natural-language-processing-nlp, Visted on 12/06/2021.

[6] Eyrolles, Apprentissage automatique, Release date: 17/05/2018.

[7] Apprentissage automatique, http://www.leparisien.fr/, Visted on 12/06/2021.

[8] Walaa Medhat, Ahmed Hassan, Hoda Korashy, Sentiment analysis algorithms and applications: A survey, Ain Shams Engineering Journal (2014) 5, 1093–1113.

[9] https://searchenterpriseai.techtarget.com/definition/machine-learning-ML, Visited on 12/06/2021.

[10] https://finnaarupnielsen.wordpress.com, Visited on 12/06/2021.

[11] http://www.wjh.harvard.edu, Visited on 12/06/2021.

[12] http://www.sentic.net, Visited on 12/06/2021.

[13] http://www.sentiwordnet.isti.cnr.it/, Visited on 12/06/2021.

[14] sentiment140. URL : http://www.sentiment140.com Visited on 12/06/2021.

[15] http://www.tweetfeel.com, Visited on 12/06/2021.

[16] Chenlo J, Hogenboom A, Losada D. Sentiment-based ranking of blog posts using rhetorical structure theory. In: Presented at the 18th international conference on applications of Natural Language to Information Systems (NLDB'13); 2013.

[17] http://twitrratr.com, Visited on 12/06/2021.

[18] http://smm.streamcrab.com, Visited on 12/06/2021.

[19] "Colaboratory Google.,". https://research.google.com/colaboratory/faq.html, Visited on 12/06/2021.

[20] I. S. a. G. E. H. A. Krizhevsky, "ImageNet Classification with Deep ConvolutionalNeural Networks," no. 1097–1105, 2012.

[21] A. Géron, Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472., 2017.

[22] "k-Nearest Neighbors,", http://www.statsoft.com/Textbook/k-Nearest-Neighbors, Visited on 12/06/2021.

[23] R. Gandhi, "Support Vector Machine — Introduction to Machine Learning Algorithms," https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47, Visited on 12/06/2021.

[24] Christenburg, "Artificial intelligence: What is it? What are the most prominent manifestations?," https://digitrate.us/artificial-intelligence/artificial-intelligence-what-is-it-what-are-the-most-prominent-manifestations/, Visited on 12/06/2021.

[25] X. C. Zhihua Cui, Swarm intelligence and bio-inspired computation: theory and applications, Elsevier, 2013.

[26] "convolutional-networks" http://cs231n.github.io/convolutional-networks/, Visited on 12/06/2021.

[27] "Dropout Layer," https://leonardoaraujosantos.gitbooks.io/artificial-inteligence/content/dropout_layer.html, Visited on 12/06/2021.

[28] https://medium.com/@margaretmz/running-jupyter-notebook-with-colab#f4a29a9c7156, Visited on 12/06/2021.

[29] Cao Qing, Duan Wenjing, Gan Qiwei. Exploring determinants of voting for the''helpfulness'' of online user reviews: a text mining approach. Decis Support Syst 2011;50:511–21. [30]

[30] Chin Chen Chien, Tseng You-De. Quality evaluation of product reviews using an information quality framework. Decis Support Syst 2011;50:755–68.

[31] Moreo A, Romero M, Castro JL, Zurita JM. Lexicon-based ,comments-oriented news sentiment analyzer system. Expert Syst Appl 2012;39:9166–80.

[32] Chenlo J, Hogenboom A, Losada D. Sentiment-based ranking of blog posts using rhetorical structure theory. In: Presented at the 18th international conference on applications of Natural Language to Information Systems (NLDB'13); 2013.

[33] https://www.techno-science.net, Visited on le 12/06/2021.