Order No :
Serial No :

People's Democratic Republic of Algeria

Ministry of Higher Education and Scientific Research

ECHAHID HAMMA LAKHDAR UNIVERSITY OF EL OUED

FACULTY OF EXACT SCIENCES

COMPUTER SCIENCE DEPARTMENT

End of study thesis

ACADEMIC MASTER

Field: Mathematics and Computer Science

Sector: Computer Science

Speciality: Distributed Systems and Artificial Intelligence (SDIA)

Theme

Machine learning for religious communities' detection on social networks.

Submitted by:

- Adaika Khaoula
- Dou Youmna

Supervised by:

Dr. M. Meftah M.C.Eddine

Board of examiners :

M. Meftah M.C.Eddine	Supervisor	Univ. of El Oued
M. Ben Ali Abdekamel	President	Univ. of El Oued
M.Soltani Khaled	Examiner	Univ. of El Oued

Academic year : 2019 – 2020.

Dedication

It is often said that the journey is as important as the destination. Five years have allowed us to understand the meaning of this very simple sentence. This journey was not accomplished without challenges and without raising many questions for which the answers led to long hours of work.

I would first like to thank God who gave me the chance, the will and the patience to realize my first dream of being successful in my studies.

To my very dear parents: My father and my mother, who instilled in me the spirit of combativeness and perseverance and who always pushed and motivated me in my studies. Without them, I certainly would not be serious at this level. i sincerely hope that my parents, who have given me moral support, can find in these lines the expression of me deep gratitude. I thank you from the bottom of my heart, for all your sacrifices, your love, your tenderness, your support and your prayers throughout my studies, To my dear sisters for their permanent encouragement, and their moral support, To my dear brothers for their support and their encouragement, To all my family for their support throughout my university career, May this work be the fulfillment of your vows so alleged, and the leak of your infallible support, Thank you for always being there for me .

my thanks extend to my promoter DR. Meftah for his patience, his availability.

Adaika Khaoula

Dedicacation

On this day, five years ago, I was waiting for the news that would give me the ticket to get to the university, and now I'm nearing the end of my university career that was full of experiences, challenges and successes. This locomotive has gone through many obstacles, but I tried to surpass it consistently with the grace of Allah Come first of all, then my family and dear teachers I extend my sincere thanks and gratitude to my dear parents for their constant invitations, support, and continuous support. They were like braces and support in order to complete my university career I ask Allah to prolong their lives. To my dear sister, the permanent bond. To my dear brothers ... for their support and encouragement, To my cousin, for his constant encouragement and constant support. To all my family for their support throughout my university career. I should not forget my teachers who had the greatest role in helping me and providing me with valuable information. I ask Allah to help them and help them in their educational career.

DOU YOUMNA

Dedication

We like today to offer our thanks to our supervisor who gave and still gives us a lot of his time thoughts and efforts with out waiting for praise or thanks.

Our teacher and supervisor is creative in providing assistance and help wich makes everyone in Hamma Lakhdar university respect him.

Our Teacher Meftah Mohammed Charaf eddine these words are only simple expressions of appreciation and respect .Thank you our teacher and May ALLAH bless you

ADAIKA KHAOULA

DOU YOUMNA

Chapter 1 Concepts on Artificial Intelligence and machine learning

Résumé:

Les réseaux sociaux représentent aujourdhui un moyen de communication incontournable.

Aujourd'hui, l'analyse des textes a une grande importance dans les domaines vitaux et quautidien comme politiques, productions et services…etc. Actuellement, les réseaux sociaux pleins des textes dans lesquelles, les internautes s'expriment en différents sujets, l'intérêt de leurs opinions est considérable, où la compréhension du contenu véhiculé par ces textes est un élément essentiel.

Nous avons utilisé l'algorithme de classification suivant : Support vector machines,Decision Tree classification,Random Forest Classification,Naïve Bayes

Dans ce travail, nous mettrons en œuvre des algorithmes prédéfinis pour analyser et classer un ensemble de commentaires tirés des réseaux sociaux.

Les catégories que nous avons identifiées sont les suivantes: Musulman, Chrétien, Juif et Athée, car nous ajoutons une autre catégorie pour les commentaires neutres.

L'objectif principal de l'étude de notre mémoire est de classer les religions des peuples à travers leurs commentaires sur les réseaux sociaux. Notre travail fait partie des travaux qui utilisent et comparent un ensemble d'algorithmes pour classer les commentaires dans les réseaux sociaux.

Mots clés : Réseaux Sociaux, Apprentissage Automatique, Traitement Automatique du Langage Naturel, Fouille de texte,

Abstract:

Today, social networks are an essential means of communication.

At present , the analysis of texts was of great importance especially in fields such as policies, productions and services... etc. At present, the social networks full of texts in which, Internet users express themselves in different subjects, the interest of their opinions is considerable, where the understanding of the content conveyed by these texts is an essential element.

We used the following classification algorithms: Support vector machines,Decision Tree classification,Random Forest Classification,Naïve Bayes

In this work we will implement pre-defined algorithms to analyze and classify a set of comments drawn from social networks.

The categories we identified are as follows: Muslim, Christian, Jewish and atheist, as we add another category for neutral comments.

The main objective of the study of our memory is to classify the religions of peoples through their comments on social networks. Our work is among the works that use and compare a set of algorithms to classify comments in social networks.

Chapter 1 Concepts on Artificial Intelligence and machine learning

Key words: Social Networks, Machine Learning, Natural Automatic Language Processing, Text Mining.

ملخص:

تمثل الشبكات الاجتماعية اليوم وسيلة أساسية للاتصال.

في الوقت الحاضر، يعتبر تحليل النصوص ذا أهمية كبيرة خاصة في هذه المجالات: السياسة ; الإنتاج والخدمات ... إلخ، الشبكات الاجتماعية االيوم مليئة بالنصوص التي يعبر فيها مستخدمو الإنترنت عن أنفسهم في مواضيع عديدة ومختلفة، حيث يعد فهم المحتوى الذي تنقله هذه النصوص عنصرًا أساسيًا.

قمنا باستخدام أربعة خوارزميات للتصنيف كالتالي:دعم الالات المتجهة, شجرة القرار,الغابات العشوائية , Naive Bayes

في هذا العمل سنقوم بتطبيق خوارزميات محددة مسبقًا لتحليل وتصنيف مجموعة من التعليقات المستمدة من الشبكات الاجتماعية.

الفئات التي حددناها هي كما يلي: مسلمون ومسيحيون ويهود وملحدون، كما اضفنا فئة أخرى للتعليقات المحايدة.

الهدف الرئيسي من دراسة مذكرتنا هو تصنيف أديان الشعوب من خلال تعليقاتهم على الشبكات الاجتماعية.
عملنا يعتبر من بين الأعمال التي توضف وتقارن مجموعة من الخوارزميات لتصنيف التعليقات في الشبكات الاجتماعية.

الكلمات المفتاحية: الشبكات الاجتماعية، التعلم الآلي، المعالجة التلقائية للغة الطبيعية، تعدين النصوص.

Chapter 1 Concepts on Artificial Intelligence and machine learning

General Introduction:

Social media is a popular way to chat, express views, share content, and promote ideas and products. However, any form of expression is today freed from the idea of border, each one having the possibility of broadcasting content to a potentially unlimited audience because of the ease and speed of use and the possibility of remaining anonymou.In 2018, more than 3 billion people (40% of the world's population) connected to social media. These mainly allow the communication of content between active users.[1]

The aim of this work is to study comments on the level of social networks. More precisely Facebook and Twitter, in order to discover the various religious societies

In our thesis, we highlighted the following religions: Islam. Christianity. Judaism and atheism.Using 4 algorithms for machine learning which are: Support vector machines, Decision Tree classification,Random Forest Classification,Naïve Bayes . we identify the features of each religion to obtain a more accurate data set and then we collected the data to which the form is applied then we built dictionaries.

In the first chapter, we try to give a definition of artificial intelligence and mention its importance. After that, we present the definition of machine learning, its different types and multiple algorithms. As for the second chapter, we devote it to the Social Network Analysis and Text Mining We also mention the methods for classifying texts and mechanisms used in that. Finally, we present Machine Learning for Natural Language Processing. In the third Chapter we present the works related to our thesis.In Chapter 4 we focus on the modeling of our system in which we present the features and dictionaries that we used in this work.

Finally, we devote the last chapter to explain the Python code used and we mention the experiments and discuss the results obtained.

Chapter 1 Concepts on Artificial Intelligence and machine learning

Table of contents

List of Figures:

List of Tables:

List of Acronyms:

AI	Artificial Intelligence.
NLP	Natural Language Processing.
API	Application Programming Interfaces.
KNN	K-Nearest Neighbors.
SVM	Support Vector Machine
DT	Decision Trees.
ML	Machine Learning.
TFIDF	Term Frequency Inverse Document Frequency.
NALP	Natural Automatique Language Processing.
AM	Morphological Analyser.
BOW	Bag -Of –Words.
CNN	Convolutional Neural Network.
POS	Parts Of Speech.
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative.

Chapter 1: Concepts on Artificial Intelligence and machine learning

1. Concepts on Artificial Intelligence and machine learning

1.1. Introduction :

With the tremendous increase in the use of social network sites like Twitter and Facebook,online community is exchanging information in the form of opinions, sentiments, emotions, and intentions, which reflect their affiliations and aptitude towards an entity, event and religions.Opinions expressed on such sites give an important clue about the activities and behavior of online users.

In this chapter, we will provide a general definition of artificial intelligence and show how it works and its subfields and we define machine learning. We will mention the data that algorithms use in making predictions and decisions. These data are in turn divided into training and test data. Then we will talk in detail about the types of machine learning.

1.2Artificial Intelligence:

1.2.1 Artificial Intelligence History:

The term artificial intelligence was coined in 1956, but AI has become more popular today thanks to increased data volumes, advanced algorithms, and improvements in computing power and storage. [2]

The 1950s explored topics like problem solving and symbolic methods. in the 1960s, the us department of defense took interest in this type of work and began training computers to mimic basic human reasoning.For example, the defense advanced research projects agency (darpa) completed street mapping projects in the 1970s. and darpa produced intelligent personal assistants in 2003, long before siri, alexa or cortana were household names.This early work paved the way for the automation and formal reasoning that we see in computers today, including decision support systems and smart search systems that can be designed to complement and augment human abilities.[3]

1.2.2 What is Artificial Intelligence?

Artificial intelligence (AI) is wide-ranging branch of computer science concerned with building smart machines capable of performing tasks that typically require human intelligence. AI is an interdisciplinary science with multiple approaches, but advancements in machine learning and deep learning are creating a paradigm shift in virtually every sector of the tech industry.[4]

1.2.3 Why is artificial intelligence important? :

- AI automates repetitive learning and discovery through data. But AI is different from hardware-driven, robotic automation. Instead of automating manual tasks, AI performs frequent, high-volume, computerized tasks reliably and without fatigue. For this type of automation, human inquiry is still essential to set up the system and ask the right questions.

- AI adds intelligence to existing products. In most cases, AI will not be sold as an individual application. Rather, products you already use will be improved with AI capabilities, much like Siri was added as a feature to a new generation of Apple products. Automation, conversational platforms.

- Bots and smart machines can be combined with large amounts of data to improve many technologies at home and in the workplace, from security intelligence to investment analysis.[2]

1.2.4 How Artificial Intelligence Works:

AI works by combining large amounts of data with fast, iterative processing and intelligent algorithms, allowing the software to learn automatically from patterns or features in the data. AI is a broad field of study that includes many theories, methods and technologies, as well as the following major subfields: [2]

- **Machine learning** :

  Automates analytical model building. It uses methods from neural networks, statistics, operations research and physics to find hidden insights in data without explicitly being programmed for where to look or what to conclude. [2]

- **Neural network :**

A neural network is a type of machine learning that is made up of interconnected units (like neurons) that processes information by responding to external inputs, relaying information between each unit. The process requires multiple passes at the data to find connections and derive meaning from undefined data.[2]

- **Deep learning :**

Uses huge neural networks with many layers of processing units, taking advantage of advances in computing power and improved training techniques to learn complex patterns in large amounts of data. Common applications include image and speech recognition.[2]

- **Cognitive Computing** :

Is a subfield of AI that strives for a natural, human-like interaction with machines. Using AI and cognitive computing, the ultimate goal is for a machine to simulate human processes through the ability to interpret images and speech – and then speak coherently in response.[2]

- **Natural Language Processing (NLP) :**

Is the ability of computers to analyze, understand and generate human language, including speech. The next stage of NLP is natural language interaction, which allows humans to communicate with computers using normal, everyday language to perform tasks.

Additionally, several technologies enable and support AI:[2]

- **Graphical processing units:**

Are key to AI because they provide the heavy compute power that's required for iterative processing. Training neural networks requires big data plus compute power. [2]

- **The Internet of Things** :

Generates massive amounts of data from connected devices, most of it unanalyzed. Automating models with AI will allow us to use more of it.[2]

1.3Machine Learning – Definition :

Machine Learning is a sub-area of artificial intelligence.whereby the term refers to the ability of information technologie systems to independently find solutions to problems by recognizing patterns in databases.

In other words, machine Learning enables IT systems to recognize patterns on the basis of existing algorithms and data sets and to develop adequate solution concepts. Therefore, in Machine Learning, artificial knowledge is generated on the basis of experience.

In order to enable the software to independently generate solutions, the prior action of people is necessary.for example, the required algorithms and data must be fed into the systems in advance and the respective analysis rules for the recognition of patterns in the data stock must be defined.

Once these two steps have been completed, the system can perform the following tasks by Machine Learning:

- Finding, extracting and summarizing relevant data
- Making predictions based on the analysis data
- Optimizing processes based on recognized patterns Advantages of Machine Learning Machine Learning undoubtedly helps people to work more creatively and efficiently. Basically, you too can delegate quite complex or monotonous work to the computer through Machine Learning.[5]

1.3.1 Training, validation, and test sets

In machine learning, a common task is the study and construction of algorithms that can learn from and make predictions on data.Such algorithms work by making data-driven predictions or decisions, through building a mathematical model from input data.

The model is initially fit on a training dataset.The model is trained on the training dataset using a supervised learning method, In practice, the training dataset often consists of pairs of an input vector and the corresponding output vector , where the

answer key is commonly denoted as the target .The current model is run with the training dataset and produces a result, which is then compared with the target. Based on the result of the comparison and the specific learning algorithm being used, the parameters of the model are adjusted. The model fitting can include both variable selection and parameter estimation.

And to develop the AI and ML model, a precise training data is required that help algorithms to understand the certain patterns or series of outcomes comes to a given question. And training data can consist texts, images or videos which are mainly labeled to make it recognizable to computer vision and understandable to machines.

In machine learning training data is the key factor to make the machines recognize the objects or certain patterns and make the right prediction when used in real-life. Basically, there are three types of training data used in machine learning model development and each data has its own importance and role in building a ML model.[6]

- **Training dataset :**

Training data is basically a type of data used for training a new application, model or system through various methods depending on the project's feasibility and requirements that is to fit the parameters . And training data for AI or ML is slightly different, as they are labeled or annotated with certain techniques to make it recognizable to computer that helps machines to understand the objects.

This data set is used by machine learning engineer to develop your algorithm and more than 70% of your total data used in the project. A huge quantity of datasets are used to train the model at best level to get the best results.[6]

- **Test dataset :**

This is the second type of data helps to check the prediction level of machine learning and AI model. Its is similar to validation data in testing the model accuracy but don't help to improve the prediction level. It is basically used to test the model whether it will work well in real-life use and final test in the moment of truth for the model, if it works

perfectly.here it is notable that all these three types of data are important for right ML model development. [6]

- **Validation dataset :**

A validation dataset is a dataset of examples used to tune the hyperparameters of a classifier. It is sometimes also called the development set or the "dev set". In artificial neural networks, a hyperparameter is,It, as well as the testing set , should follow the same probability distribution as the training dataset.

In order to avoid overfitting, when any classification parameter needs to be adjusted, it is necessary to have a validation dataset in addition to the training and test datasets. For example, if the most suitable classifier for the problem is sought, the training dataset is used to train the candidate algorithms, the validation dataset is used to compare their performances and decide which one to take and, finally, the test dataset is used to obtain the performance characteristics such as accuracy, sensitivity, specificity, F-measure, and so on. The validation dataset functions as a hybrid : it is training data used for testing, but neither as part of the low-level training nor as part of the final testing.[6]

1.3.2 Methods of Machine Learning:

In Machine Learning, statistical and mathematical methods are used to learn from data sets. Dozens of different methods exist for this, whereby a general distinction can be made between two systems, namely symbolic approaches on the one hand and sub-symbolic approaches on the other. While symbolic systems are, for example, propositional systems in which the knowledge content, i.e.The induced rules and the examples are explicitly represented , sub-symbolic systems are artificial neuronal networks.These work on the principle of the human brain, whereby the knowledge contents are implicitly represented.[5]

1.3.3 Types of Machine Learning:

Basically, Algorithms play an important role in Machine Learning: On the one hand, they are responsible for recognizing patterns and on the other hand, they can generate solutions. [5]

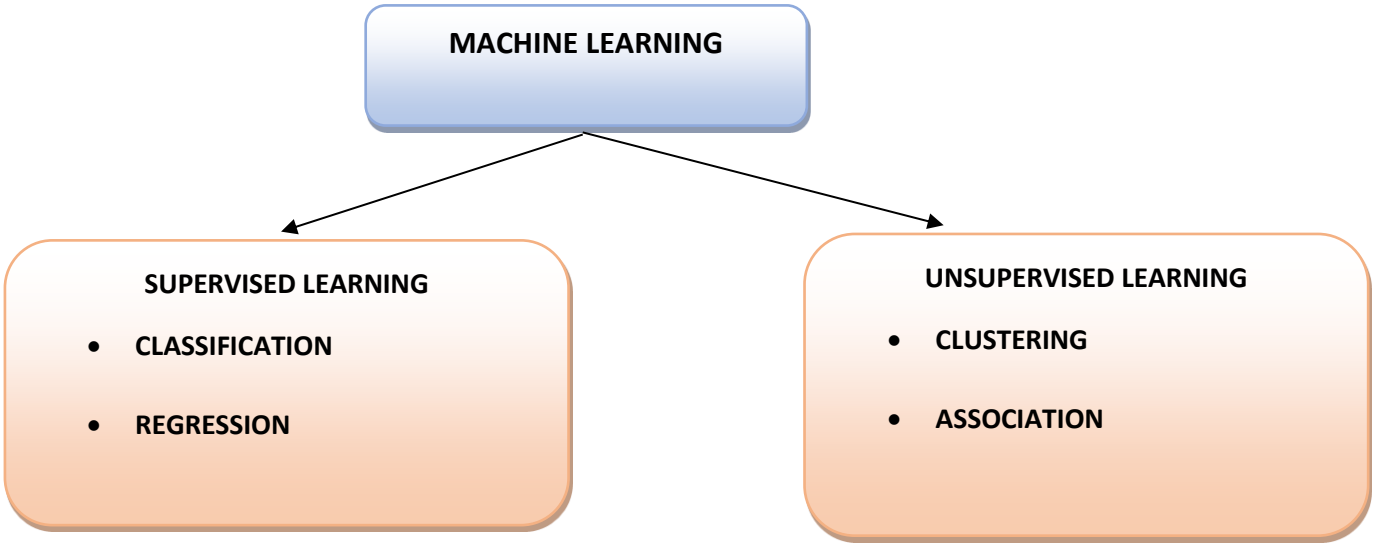


Figure 1.1: Types of Machine Learning [6]

1.3.3.1 Supervised learning:

In the course of monitored learning, example models are defined in advance. In order to ensure an adequate allocation of the information to the respective model groups of the algorithms, these then have to be specified. In other words, the system learns on the basis of given input and output pairs. In the course of monitored learning, a programmer, who acts as a kind of teacher, provides the appropriate values for a particular input. The aim is to train the system in the context of successive calculations with different inputs and outputs and to establish connections.[5]

13.3.1.1 Classification

In these types of problems, we predict the response as specific classes, such as “yes” or “no”. When only 2 classes are present, then it is called a Binary Classification. For more than 2 class values, it is called a Multi-class Classification. The predicted response values are discrete values. For example, Is it the image of the sun or the moon? The classification algorithm separates the data into classes.[8]

13.3.1.2 Regression:

Regression problems predict the response as continuous values such as predicting a value that ranges from -infinity to infinity. It may take many values. For example, the linear regression algorithm that is applied, predicts the cost of the house based on many parameters such as location, nearby airport, size of the house, etc.[8]

1.3.3.2 Unsupervised learning:

In unsupervised learning, artificial intelligence learns without predefined target values and without rewards. It is mainly used for learning segmentation (clustering). The machine tries to structure and sort the data entered according to certain characteristics. For example, a machine could (very simply) learn that coins of different colours can be sorted according to the characteristic "colour" in order to structure them.[5]

1.3.3.2.1 Clustering

Is a technique of organising a group of data into classes and clusters where the objects reside inside a cluster will have high similarity and the objects of two clusters would be dissimilar to each other. Here the two clusters can be considered as disjoint. The main target of clustering is to divide the whole data into multiple clusters. Unlike classification process, here the class labels of objects are not known before, and clustering pertains to unsupervised learning.

1.3.3.2.2 Association

Rules allow you to establish associations amongst data objects inside large databases. This unsupervised technique is about discovering interesting relationships between variables in large databases. For example, people that buy a new home most likely to buy new furniture.[9]

1.3.4 Machine Learning: The technology leaders :

In addition to Microsoft, Google, Facebook, IBM and Amazon, Apple also spends enormous financial resources on the use and further development of Machine Learning. IBM's Watson supercomputer is still the best-known appliance for Machine Learning.Watson is mainly used in the medical and financial sectors. As already mentioned, Facebook uses Machine Learning for image recognition, Microsoft for the

speech recognition system Cortana, Apple for Siri. Of course, Machine Learning is also used at Google, both in the area of image services and search engine ranking.Cloud providers such as Google, Microsoft, Amazon Webservice and IBM have now created services for Machine Learning. With their help it is also possible for developers who do not have specific Machine Learning knowledge to develop applications. These applications are able to learn from a freely definable set of data. Depending on the provider, these platforms have different names:

- IBM: Watson
- Amazon: Amazon Machine Learning
- Microsoft: Azure ML Studio
- Google: Tensorflow .[5]

1.3.5 Classification - Machine Learning:

Classification is a process of categorizing a given set of data into classes, It can be performed on both structured or unstructured data. The process starts with predicting the class of given data points. The classes are often referred to as target, label or categories.

The classification predictive modeling is the task of approximating the mapping function from input variables to discrete output variables. The main goal is to identify which class/category the new data will fall into.

Let us look at some of the objectives covered under this section of our thesis.

- Define Classification and list its algorithms
- Describe Logistic Regression and Sigmoid Probability
- Explain K-Nearest Neighbors and KNN classification Understand Support Vector Machines.
- Implement the Naïve Bayes Classifier
- Demonstrate Decision Tree Classifier
- Describe Random Forest Classifier. [10]

1.3.5.1 Algorithms of Classification:

Let's have a quick look into the types of Classification Algorithm below.

- Linear Models:
  - Logistic Regression
  - Support Vector Machines
- Nonlinear models:
  - K-nearest Neighbors (KNN)
  - Naïve Bayes
  - Decision Tree Classification
  - Random Forest Classification [11]

1.3.5.1.1 Logistic Regression: Meaning

Let us understand the Logistic Regression model below.

- This refers to a regression model that is used for classification.
- This method is widely used for binary classification problems. It can also be extended to multi-class classification problems.Here, the dependent variable is categorical: $y \epsilon \{0, 1\}$
- A binary dependent variable can have only two values, like 0 or 1, win or lose, pass or fail, healthy or sick, etc
- In this case, you model the probability distribution of output y as 1 or 0. This is called the sigmoid probability (σ).
- If $\sigma(\theta Tx) > 0.5$, set $y = 1$, else set $y = 0$
- Unlike Linear Regression (and its Normal Equation solution), there is no closed form solution for finding optimal weights of Logistic Regression. Instead, you must solve this with maximum likelihood estimation (a probability model to detect the maximum likelihood of something happening).
- It can be used to calculate the probability of a given outcome in a binary model, like the probability of being classified as sick or passing an exam.

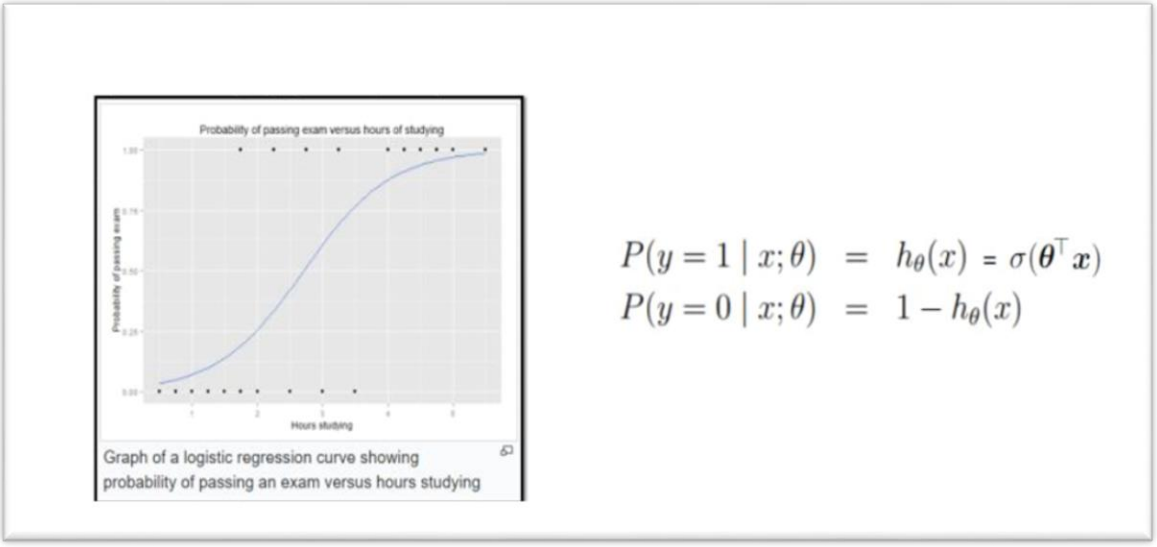


Figure 1.2: Graph of logistic regression curve showing probability of passing an exam versus hours studying [11]

Sigmoid Probability

- The probability in the logistic regression is often represented by the Sigmoid function (also called the logistic function or the S-curve):

S(T) = $\frac{1}{1+e^{-t}}$

- In this equation, t represents data values * the number of hours studied and S(t) represents the probability of passing the exam.
- Assume sigmoid function:

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

$$g(z) = \frac{1}{1 + e^{-z}}$$

- g(z) tends toward 1 as z -> infinity , and g(z) tends toward 0 as z -> infinity [11]

1.3.5.1.2 Support Vector Machine (SVM)

Let us understand Support Vector Machine (SVM) in detail below.

- SVMs are classification algorithms used to assign data to various classes.

- They involve detecting hyperplanes which segregate data into classes.
- SVMs are very versatile and are also capable of performing linear or nonlinear classification, regression, and outlier detection.
- Once ideal hyperplanes are discovered, new data points can be easily classified.

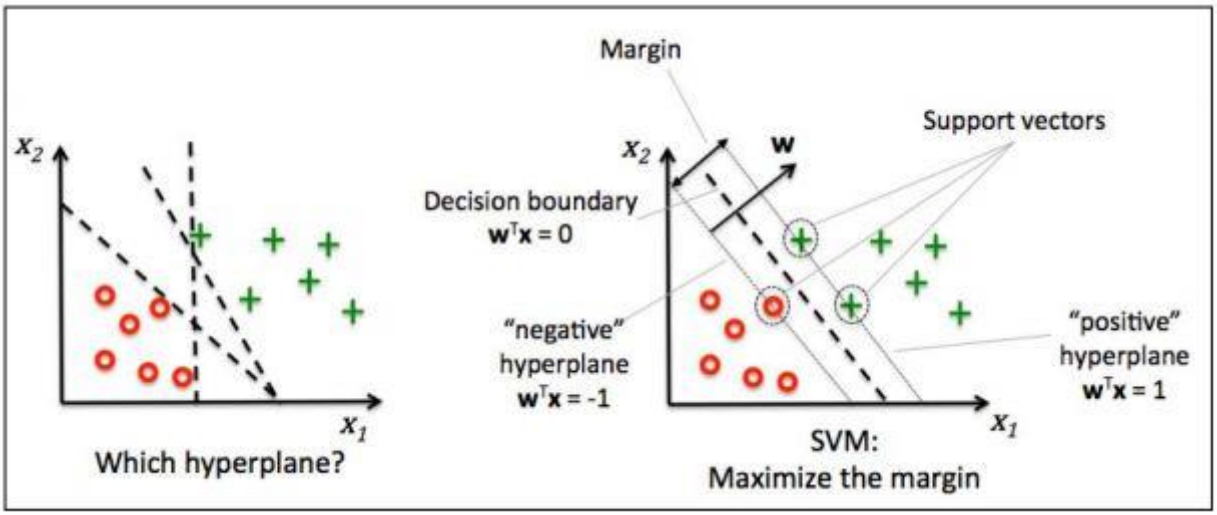


Figure 1.3: Exemple of Support Vector Machine [11]

- The optimization objective is to find "maximum margin hyperplane" that is farthest from the closest points in the two classes (these points are called support vectors).
- In the given figure, the middle line represents the hyperplane.[11]

1.3.5.1.3 k-nearest Neighbors (KNN)

K-nearest Neighbors algorithm is used to assign a data point to clusters based on similarity measurement. It uses a supervised method for classification.

The steps to writing a k-means algorithm are as given below:

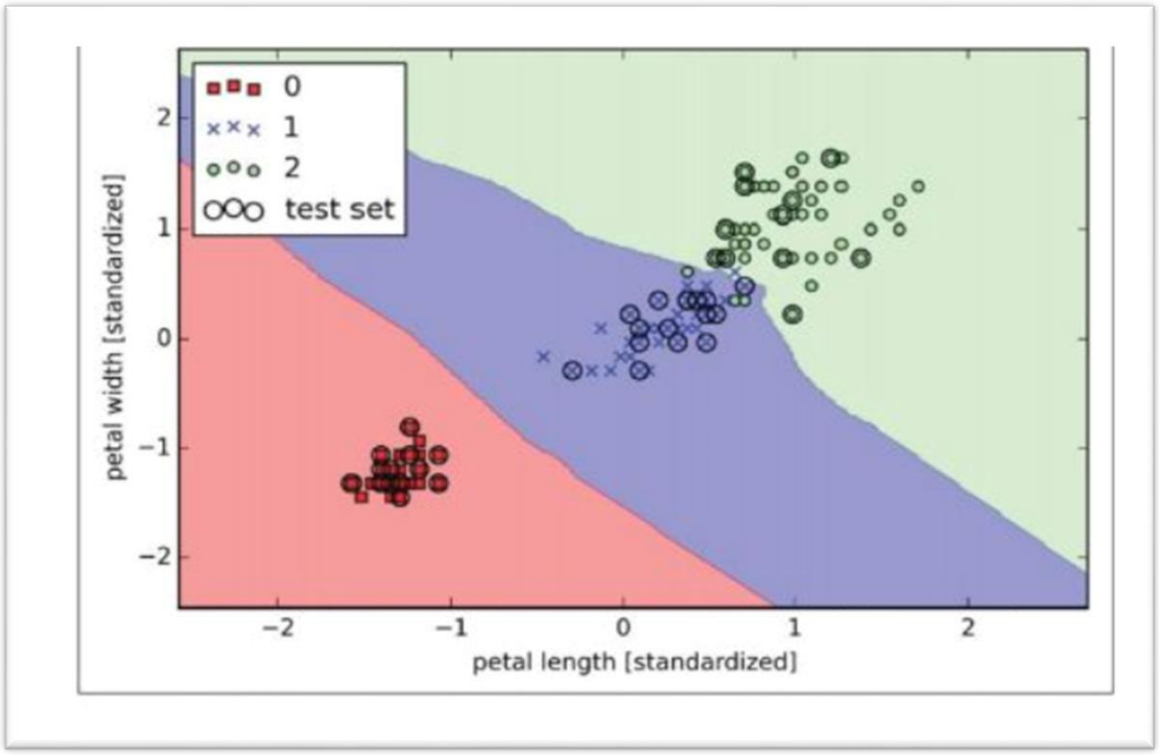


Figure 1.4: Exemple of K-nearest Neighbors. [11]

- Choose the number of k and a distance metric. (k = 5 is common)
- Find k-nearest neighbors of the sample that you want to classify
- Assign the class label by majority vote. [11]

1.3.5.1.4 KNN Classification

A new input point is classified in the category such that it has the most number of neighbors from that category. For example:

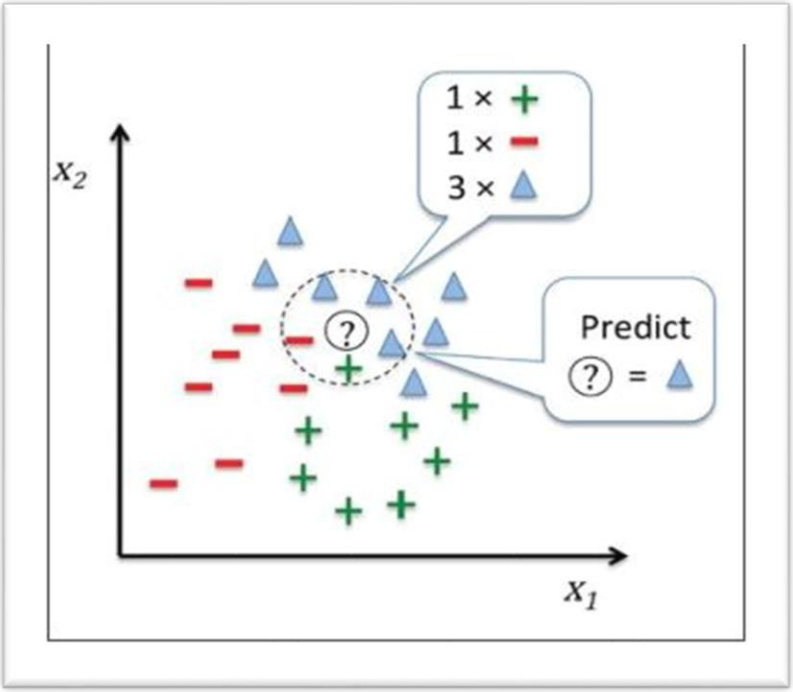


Figure 1.5: Exemple of KNN classification [11]

- A patient as high risk or low risk.
- Mark email as spam or ham. [11]

1.3.5.1.5 Naive Bayes

Naive Bayes classifier works on the principle of conditional probability as given by the Bayes theorem. [11]

1.3.5.1.5.1 Bayes Theorem:

We will understand Bayes Theorem in detail from the points mentioned below.

According to the Bayes model, the conditional probability P(Y|X) can be calculated as:

P(Y|X) = P(X|Y)P(Y) / P(X)

This means you have to estimate a very large number of P(X|Y) probabilities for a relatively small vector space X. [10]

For example, for a Boolean Y and 30 possible Boolean attributes in the X vector, you will have to estimate 3 billion probabilities P(X|Y).

To make it practical, a Naïve Bayes classifier is used, which assumes conditional independence of P(X) to each other, with a given value of Y.

This reduces the number of probability estimates to 2*30=60 in the above example.

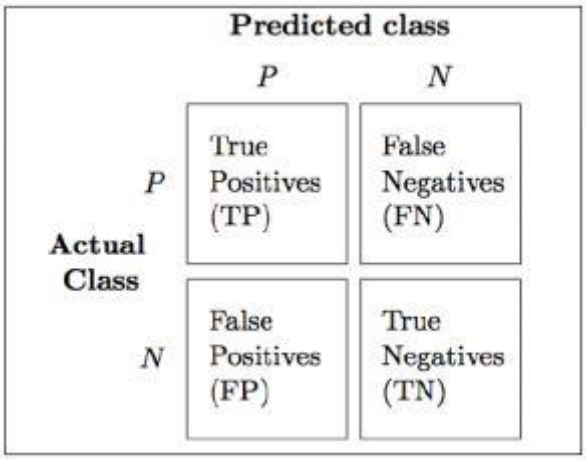


Figure 1.6: Actual class [10]

Although confusion Matrix is useful, some more precise metrics are provided by Precision and Recall. Precision refers to the accuracy of positive predictions.

$$\text{Precision}=\frac{TP}{TP+FP}$$

Recall refers to the ratio of positive instances that are correctly detected by the classifier (also known as True positive rate or TPR).[11]

$$\text{Recall}=\frac{TP}{TP+FN}$$

1.3.5.1.6 Decision Tree Classifier

Some aspects of the Decision Tree Classifier mentioned below are.

- Decision Trees (DT) can be used both for classification and regression.
- The advantage of decision trees is that they require very little data preparation.
- They do not require feature scaling or centering at all.
- They are also the fundamental components of Random Forests, one of the most powerful ML algorithms.

- Unlike Random Forests and Neural Networks (which do black-box modeling), Decision Trees are white box models, which means that inner workings of these models are clearly understood.
- In the case of classification, the data is segregated based on a series of questions.
- Any new data point is assigned to the selected leaf node.

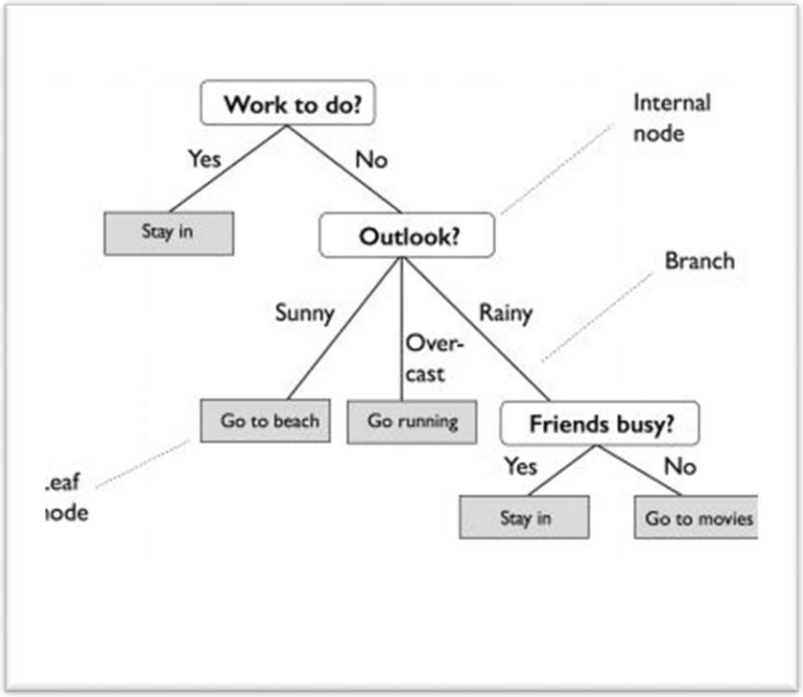


Figure 1.7: Exemple of Decision Tree Classifier [11]

- Start at the tree root and split the data on the feature using the decision algorithm, resulting in the largest information gain (IG).
- This splitting procedure is then repeated in an iterative process at each child node until the leaves are pure.
- This means that the samples at each node belonging to the same class.
- In practice, you can set a limit on the depth of the tree to prevent overfitting.
- The purity is compromised here as the final leaves may still have some impurity.
- The figure shows the classification of the Iris dataset.[11]

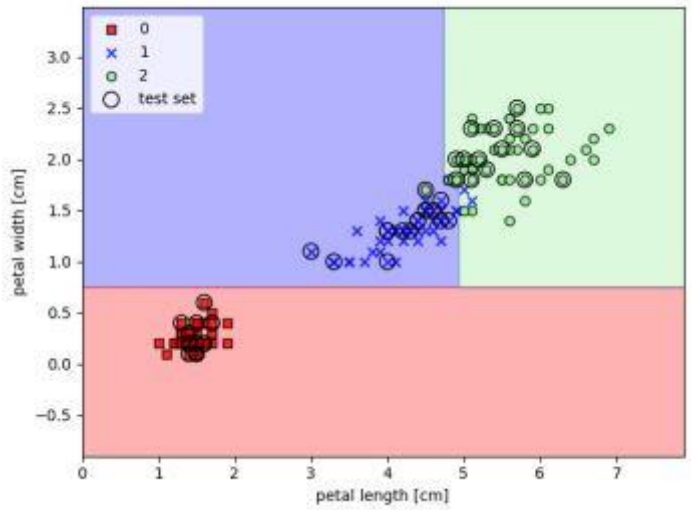


Figure 1.8: Data separation [11]

1.3.5.1.7 Random Forest Classifier

Let us have an understanding of Random Forest Classifier below.

- A random forest can be considered an ensemble of decision trees (Ensemble learning).
- Random Forest algorithm:
    - Draw a random bootstrap sample of size n (randomly choose n samples from the training set).
    - Grow a decision tree from the bootstrap sample. At each node, randomly select d features.
    - Split the node using the feature that provides the best split according to the objective function, for instance by maximizing the information gain.
    - Repeat the steps 1 to 2 k times. (k is the number of trees you want to create, using a subset of samples)
    - Aggregate the prediction by each tree for a new data point to assign the class label by majority vote (pick the group selected by the most number of trees and assign new data point to that group).
- Random Forests are opaque, which means it is difficult to visualize their inner workings.

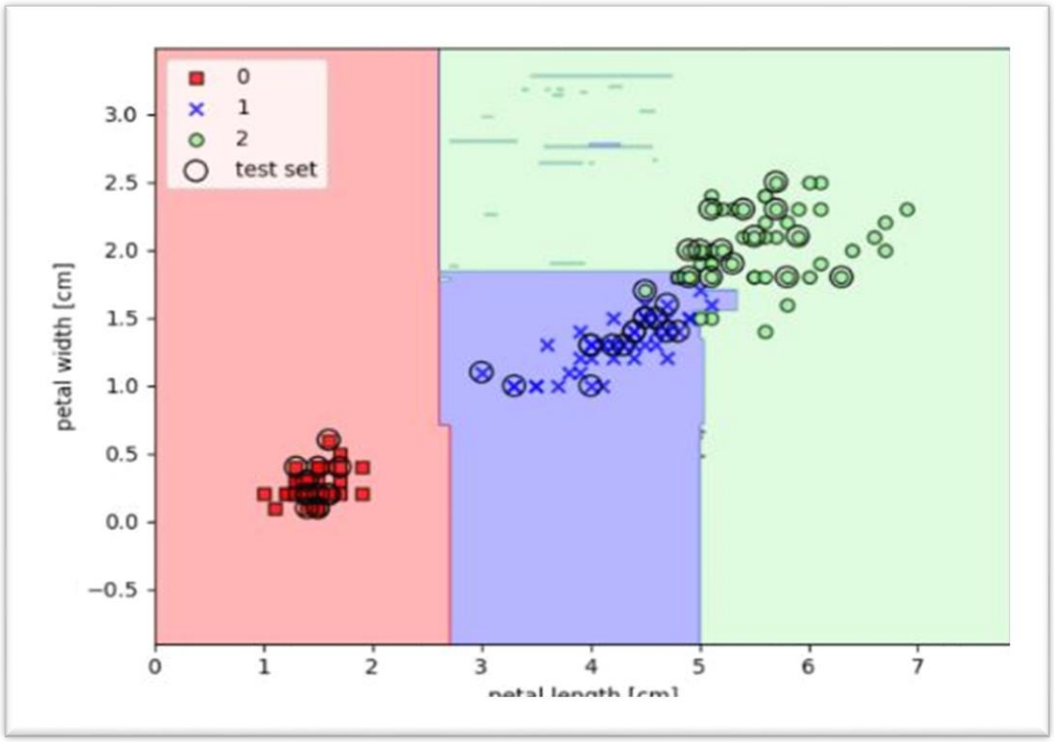


Figure 1.9: Exemple of Random Forest Classifier [11]

- However, the advantages outweigh their limitations since you do not have to worry about hyperparameters except k, which stands for the number of decision trees to be created from a subset of samples.
- RF is quite robust to noise from the individual decision trees. Hence, you need not prune individual decision trees.
- The larger the number of decision trees, the more accurate the Random Forest prediction is. (This, however, comes with higher computation cost). [11]

1.4Conclusion:

In this chapter, which is entitled "State of the Arts", we studied the different definitions that we need in our thesis. Where we provided a comprehensive definition of artificial intelligence and how it works, and we mentioned its main fields after that, we defined machine learning and mentioned the most important method that it uses, including supervised learning and unsupervised learning . Finally, we mentioned the most famous classification algorithms.In the next chapter we devote it to social networks and text mining.

Chapter 2: Social Network Analysis and Text Mining

2. Social Network Analysis and Text Mining

2.1 Introduction:

This chapter contains two branches; the first section is for social networks, where we will present the Online Social Network, their definition, their interest. after that we will present the context of analysis of social networks, for that we start by defining the term of social network which will allow us thereafter to introduce the concept of analysis of social networks as well as the methods of analysis used then We will mention the impact of Artificial Intelligence on Social media. and the second section is for NLP and Text Mining as we will mention how to classify the texts and the methods used in that.

2.2 General information on social networks:

Social networks have been ubiquitous since the advent of the Internet. They allow different users to interact in community and to group together according to criteria that are important to them. These social networks are of different types. Some are known to everyone (facebook, twitter, LinkedIn) and have millions of members. Others exploit lesser-known niches and may go relatively unnoticed or remain confidential, such as corporate networks. [12]

Brief history:

Online social networks appeared in 2002 with the American site Friendster, the first to use the principle of the online circle of friends. Then came Myspace , created in 2003.

Facebook was born in 2004. This platform created by Mark Zuckerberg was originally intended for Harvard students wishing to communicate with each other. Faced with its dazzling success, the site became mainstream in 2006. The case of Facebook marks a turning point in the democratization of social networks on the Internet.

Twitter appeared in 2006. This platform is based on the principle of microblogging: messages posted by users are limited to 140 characters. Very popular in the United States where many personalities have an account.

Initially it was a communication tool, social networks have become information dissemination tools, which, as we will see, have changed our relationship to information

and the way we discover and share it. All users of social networks are now potential producers and disseminators of information.[12]

Definition

The concept of "social" was invented in 1954 by an anthropologist named.John A. Barnes The network principle is defined by two elements: contacts and links between contacts. The more contacts we have the more important our network is, and therefore the more useful we are. On the other hand, social networks are considered to be web services that allow individuals to build a public or semi-public profile linked with a list of other profiles.

We can say that the social network can be seen as a virtual company, or the user can have an identity by creating their own profile. It also makes it possible to weave links with other members, by posting messages, articles, Photographs, etc.After having introduced the concept of social network, we present in the following the types of social networks currently existing on the internet, with a focus later on the most active networks.[13]

Purpose of social networks

Social networks meet the needs of recognition and belonging, they allow to put in contact with other people who have points or links in common, they also allow to communicate with friends, to make contact with old acquaintances lost to follow, to follow the news… etc. [13]

Categories of Social Networks on the Internet:

Here is a list of types of social networks with examples

• General social networks to discuss: they are numerous and diverse but here are the main ones to remember: Facebook, Myspace, Twitter.

• Social sharing networks: they allow uploading and sharing of videos, photos ... such as YouTube, Flicker.

- Professional social networks: they allow you to present your professional career: LinkedIn, Viadeo.

- Service networks: they offer to exchange good addresses, services, etc. [13]

Presentation of the most used social networks:

- **Facebook :**

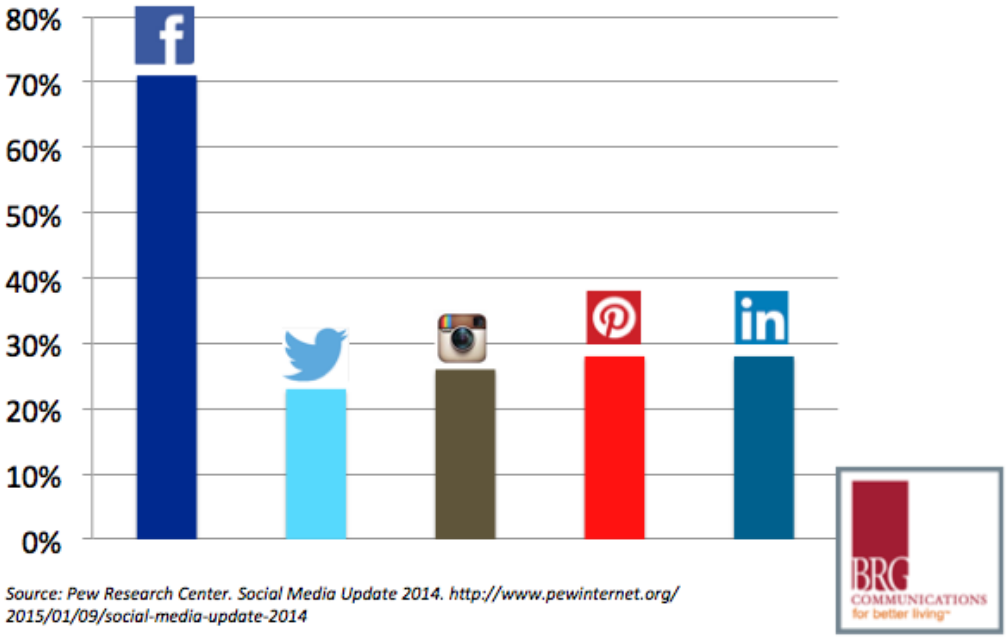


Figure 2.1: The most used social media on the Internet [14]

Created in 2006, Facebook is a social network allowing to share information, photos, videos, moods, music with friends or a community. It is possible to use Facebook privately or professionally.Each post on the wall will appear on the friends' home page. Facebook also offers an email and chat service (instant chat).[13]

Facebook continues to lead the charge as the most-used social network (71%), though user adoption has not expanded since last year. Pinterest and LinkedIn tie at second place with 28% of Internet users frequenting the platforms.[14]

- **Twitter :**

Microblogging site allowing to publish short messages of a maximum of 140 characters called "twitte". Within 140 characters, it is possible to create links to other sources and initiate a discussion with other users.

Each user can choose to follow the messages of one or more people or organizations. To make the search easier, words can be tagged this way:

#research, #education .. .[13]

- **Youtube:**

YouTube is a video hosting website where users can send,view and share video footage. It was created in February 2005 by three former PayPal employees. The service located in San Bruno in California (United States) uses the Adobe Flash and / or HTML 5 technique to display all kinds of videos: film extracts, TV shows and music clips, but also videos amateurs from blogs for example. In October 2006, Google announced that after entering into an agreement, it would become the owner of the company in exchange for Google shares worth a total of US $ 1.65 billion. The transaction ended on November 13, 2006. In 2009, 350 million people visited this video-sharing site each month. On May 17, 2010, more than 2 billionvideos were viewed daily. [13]

Analysis of social networks :

✓ Definition Analysis of social networks :

The analysis of social networks (ASN) is above all a toolbox allowing to Visualize and model social relations in the form of graphs.

- The nodes represent individuals, organizations.

- The links are the relationships between these nodes.

As a result, ASN is based on graphical visualizations resulting from algorithms allowing to calculate degrees of force or density between the different actors of a network. So the analysis of social networks is based on a structural approach to relationships between members of an organized social environment, it endeavors to describe the interdependencies between actors.

We can therefore say that the ARS is the study of social entities that is to say the people (actors) in an organization or society and their interactions (relationships).

These interactions can be represented by a graph, or each node represents an actor and each link is considered as a relation or arc.it is actually a mathematical technique that will allow us to understand the relationships that individuals establish between them through the study of the intensity of their interactions, both in the world of work and in their social community.

- A sociogram is a diagram that illustrates a social network. It can also be known as a social network diagram or graph. You will find an example of a sociogram in Figure. The circles represent different actors in the network, and the lines and arrows represent the links between them.

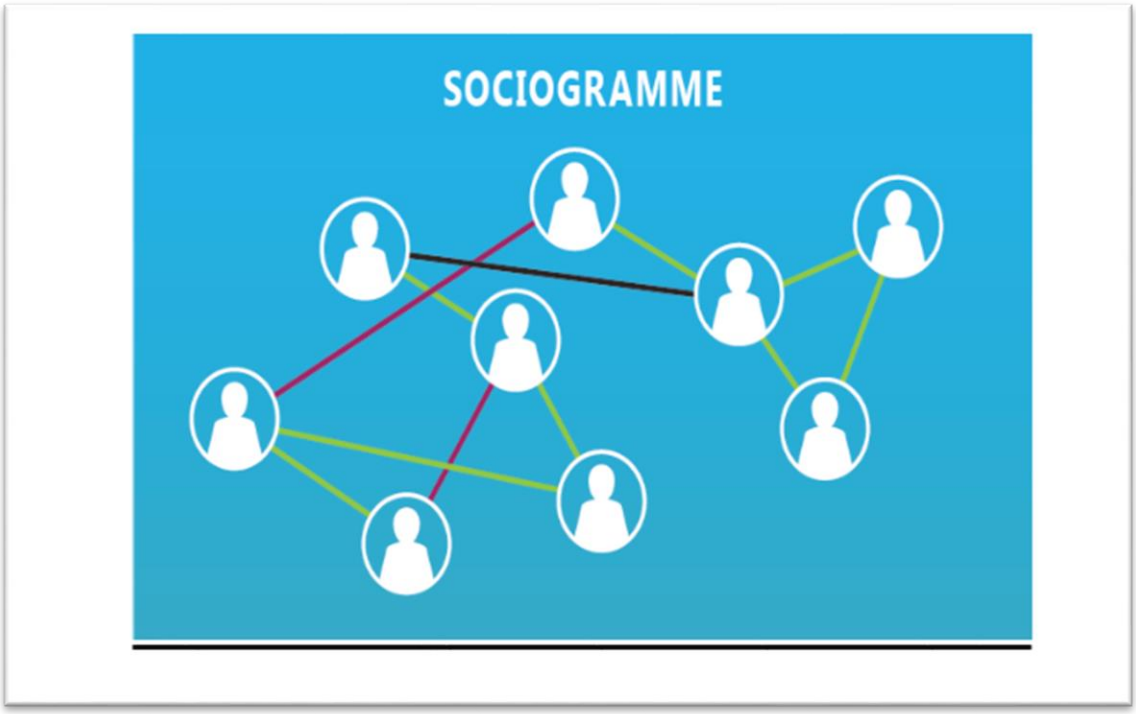


Figure 2.2: Example of a sociogram [13]

ACTORS : The circles represent individuals or organizations in the network.

LINKS : the links represent the relationships between the actors, the arrows illustrate the direction of the relationship. [13]

✓ Historical development :

ARS has almost 70 years of history. It can be divided into three main periods: the foundations of the approach; the articulation of the method; current development. The foundations of these various buildings were built between the 1940s and the 1960s.

In the 1960s and 1970s, methodological research was developed to ensure rigorous implementation. From the 1980s to today, they have been amended and perfected, which leads to the discovery of new avenues. The analysis of networks is based on two theoretical and methodological frameworks.

- The theoretical aspect is based on a broad conception of the social structure and numerous empirical studies demonstrating that the behaviors of individuals are linked to the structures in which they are inserted.

- The methodological aspect refers to the use made by the researcher of quantitative and qualitative types of data and analytical processing of these data in 1957.

Social networks are becoming more and more important in our daily life, the idea of analyzing them to extract information undoubtedly offers an important advantage for several areas.[15]

✓ Uses of social network analysis :

The large amount of information stored in social networks can be very useful in many fields.they can be used to monitor the brand of his company, by offering him information on: the opinions of the public on his product, to have an idea on the current state of the market, the competitive companies, thus to acquire a new marketing strategy because thanks to its data, it will finally be able to compare existing products on the market, ensure better management of the product range, allow better support for its customers, and be able to follow influencers .This data can also be used during elections to see public opinion.

In the rest of this part we present the methods used to analyze a social network, then establish a comparison between its methods whose purpose is to bring out the drawbacks of each and finally determine which one will be best suited to analyze today's social networks. [16]

✓ Methods of analysis of social networks:

All of these social networks collect a great deal of data on friends, messages, images, frequency of use, etc. All these exchanges and information are carefully recorded. This raises the problem of exploiting this mass of information. We must first model the network in mathematical form. The basic structure is of course the graph: the analysis of the figures produced makes it possible to draw a large amount of information, and also to predict in part the future evolution of the network.

Updates are made to offer flexible analysis methods that take into account all information in terms of structure and content. We then distinguish two large families: traditional and data mining. [17]

a) Traditional (classic) methods :

In this part, we detail the measures used in the traditional analysis of social networks. Local measures are the indices which focus on the local information of a given actor, on the other hand Global measures provide information on the entire structure (network). [18]

- **Local measures :**

The so-called local measures characterize a vertex or a link: we obtain as many results as there are vertices or links with local measures. In this section, we will introduce two concepts related to the analysis of hyperlinks which allow to measure the degree of relevance of an actor in his community : centrality and prestige.[17]

- **Centrality :**

The important actors are those who are linked and involved with other actors in an extensive way. After having modeled the contacts by links in a network, we can define a central actor as being that which is involved in several links.

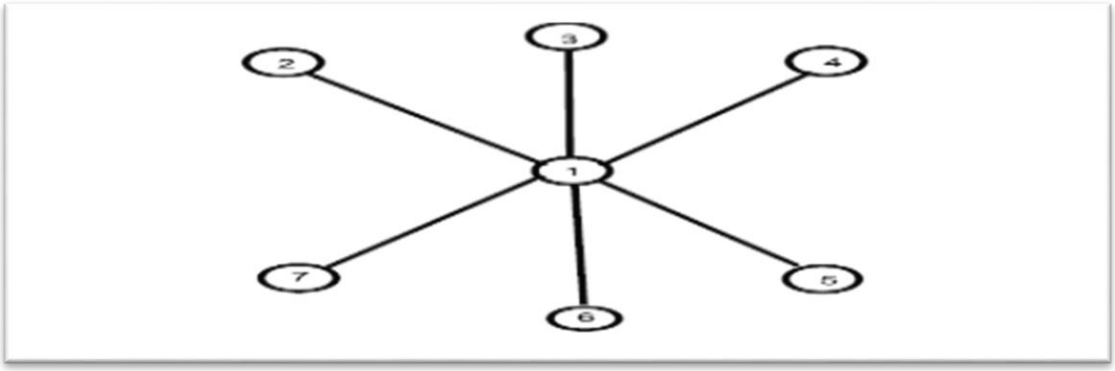


Figure 2.3: Social network center [16]

We note in the Figure that actor 1 is the most central actor because he / she communicates with the majority of the other actors, he appears on all the 6 shortest paths which link the 6 other actors.

There are different types of links between the nodes (actor), which allows to generate several types of centralities including:

- Degree of Centrality: Let n be the total number of nodes (actors) in a network (graph).

The degree of centrality of an actor i denoted CD (i) is the degree of the actor node (the number of edges) denoted d (i) normalized by the maximum degree n-1. CD(i)= d(i) / (n-l) [13]

- **Proximity centrality**:

Centrality is defined in this approach by the notion of proximity or distance. If i have a central actor then it can easily interact with the other actors. Consequently, his distance from others must be short. We therefore use the shortest distance to calculate this measurement. [13]

- **Prestige:**

The notion of prestige is another way of measuring the importance of a knot.

A prestigious node is a node to which a large number of other nodes bind - it receives a large number of inbound links. We distinguish the following measures:

- The degree of prestige: with dI (i) the incoming degree of node i, n the number of nodes in the network, the degree of prestige is given by the relation:

PD (i) = dl (i) / (n-1)

- Sorting prestige: The measures proposed so far are based on the incoming and outgoing links of a given actor.Measuring the prestige of sorting considers the reputation and importance of the actors choosing actor i. This type of algorithm is used to sort search results like Google (Rank Page algorithm).[13]

- **Global measures :**

In order to get an idea of the overall structure of the network, some global measures have been developed, we cite:

- The density P of the graph G: this measurement makes it possible to express the degree of connectivity within the graph G representing a network.It is the number of links existing in the graph G, normalized by the maximum number of links in the graph.

- The geodesic distance between two nodes is the shortest path between the two nodes.

- The average distance of a connected graph is equal to the average of the geodesic distances between all the actor pairs.

- The diameter of a connected graph is the maximum eccentricity that can exist between two of its nodes.[18]

b) Data mining in social networks :

Social networks collect a lot of data: friends, messages, images, frequency of use ... etc. This raises the problem of exploiting this mass of information. Data mining turns out to be an incredibly rich and powerful tool when applied to social networks,since this method allows us to model the network in mathematical form, analyze it to extract the maximum amount of information but also partly predict the future development of the network.[13]

- **Definition** :

Data mining or Datamining is the set of scientific methods intended for the exploration and analysis of large amounts of computer data in order to detect typical profiles, recurring behaviors, rules, links, unknown trends, specific structures that concise most of the information useful for decision support. Apply this technique in our case, this technique will allow us to extract from social networks; customer data, then analyze and classify it. Datamining allows you to perform the following four types of analysis: Classification, Estimation, Segmentation, Forecasting . These types of analysis fall into two descriptive and predictive categories as illustrated in the following table:

Desciptive technique	Predictive technique
• Classification	• Estimation • Segmentation • Forecasting

Figure 2.4: Type of data mining analysis [13]

Descriptive techniques (eg classification) make it possible to describe, summarize, synthesize and classify these techniques trying to highlight information present but hidden by the volume of data. On the other hand, predictive techniques try to extrapolate new information from the data presented above : [19]

- **Classification (descriptive technique) :**

Is a method that allows you to group objects (person, interests ... etc.) into groups, or families so that the objects of the same group are as similar as possible, and those of distinct groups differ as much as possible.

The number of groups is sometimes fixed, they are not predefined but determined during the operation. In the case of data extraction from social networks, this method will allow us to group the interests of a person by class (character, product, purchase, consumption, activity ... etc.). In another way we can say that this descriptive method allows to describe in a simple way a complex reality by summarizing it.[19]

- **Estimation (predictive technique) :**

Estimation consists in defining the link between a set of predictors and a target variable.

This link is defined from complete data, that is to say whose values are known for both the predictors and the target variable. Then we can deduce an unknown target variable from the knowledge of the predictors. The interest of this technique is to estimate from the characteristics of an object, the value of an unknown field. The estimation will be useful to detect the psychology of the person, that is to say to estimate the character of the person from these activities and environment. [19]

- **Segmentation (predictive technique) :**

Consists of forming homogeneous groups (segments) within a population. This task is often performed before the previous ones to build groups on which classification or estimation tasks are applied. In the analysis of social networks, we can say that this technique will serve us for example for the detection of communities.[19]

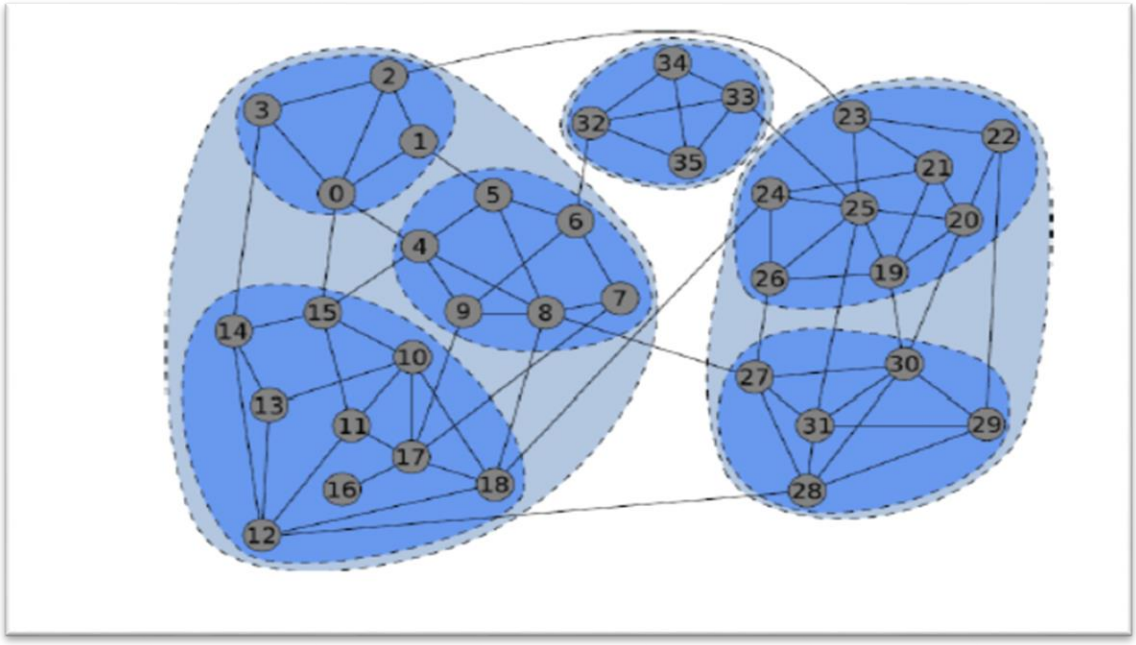


Figure 2.5: Example of segmentation [19]

- **Forecasting (predictive technique):**

Consists of estimating a future value of a field from the actual data possessed. Classification and estimation methods can be used in prediction. In our case, the forecast will serve us to prevent the behavior of a person in the future, the products that will interest him in the future. This technique is widely used in the field of artificial intelligence, especially in machine learning algorithms. [19]

The impact of Artificial Intelligence on Social media :

Social media platforms like Facebook and LinkedIn must use AI to make sense of the sea of human data coming at them.

With an estimated 2.77 billion social network users around the globe, social networking platforms possess an unimaginable amount of data. The reach of social networks is expanding at an alarming rate around the globe in 2019. These figures will continue to grow. Social media provides aphenomenal amount of user-generated data.

This is where the role of Artificial Intelligence (AI) comes into play. AI is an area of computer science that aims to decipher data from the natural world often using cognitive technologies designed to understand and complete tasks that humans have taken on in the past. [20]

The Role of AI in Social Media:

It is an era of social media.Whether it is about reaching out to new clients or nurturing existing business relationships, social media has become an integral part of interacting with businesses, influencers or customers. Effective social media management helps enhance brand's strength and increase social media conversations. It is an efficient way of engaging millions of social media users. AI helps analyze voluminous data to identify trending topics, hashtags and patterns to understand user behavior.
AI algorithms can monitor millions of unstructured user comments to understand crisis situations or trends to provide a personalized experience. With effective segmentation, AI can help provide content based on online activity and demographics. Many social networking sites have acquired AI businesses to move to the next level.

A variety of companies that offer online marketing services are exploring new ways to use social media with AI as well. They have started using AI to identify new demographics to target based on previous conversions. These AI tools rely on predictive analytics algorithms, which can extropolate information on all known users in a given social network. [20]

2.3 What is Text Mining?

Widely used in knowledge-driven organizations, text mining is the process of examining large collections of documents to discover new information or help answer specific research questions.

Text mining identifies facts, relationships and assertions that would otherwise remain buried in the mass of textual big data. Once extracted, this information is converted into a structured form that can be further analyzed, or presented directly using clustered HTML tables, mind maps, charts, etc. Text mining employs a variety of methodologies to process the text, one of the most important of these being Natural Language Processing (NLP).

The structured data created by text mining can be integrated into databases, data warehouses or business intelligence dashboards and used for descriptive, prescriptive or predictive analytics.[21]

2.4 What is Natural Language Processing (NLP)?

Natural Language Understanding helps machines “read” text (or another input such as speech) by simulating the human ability to understand a natural language such as English, Spanish or Chinese. Natural Language Processing includes both Natural Language Understanding and Natural Language Generation, which simulates the human ability to create natural language text e.g. to summarize information or take part in a dialogue.As a technology, natural language processing has come of age over the past ten years, with products such as Siri, Alexa and Google's voice search employing NLP to understand and respond to user requests.

Today’s natural language processing systems can analyze unlimited amounts of text-based data without fatigue and in a consistent, unbiased manner. They can understand concepts within complex contexts, and decipher ambiguities of language to extract key facts and relationships.[21]

2.5 The categorization of the text:

Nowadays the categorization of texts is a discipline at the crossroads of machine learning and information research and shares some of the characteristics with other tasks such as information/knowledge from texts and text search (Text Mining).The main purpose of categorizing texts is the classification of documents into a fixed number of predetermined categories. Each document will be in multiple categories, or in one, or in none.Using machine learning, the main goal is to learn from classifiers from examples that automatically assign categories Automatic text classifcation is currently used in various types of tasks, such as document reading by importance, organizing documents, automatically creating metadata, analyzing the ambiguity of words and indexing in order of documents according to the vocabulary used. In addition to being able to use a text-searching technique itself, Automatic text classification may prove to be a step in more complex text search techniques.[22]

2.6 How to categorize a text:

The categorization process incorporates the construction of a prediction model that, as a starter, receives a text and, out, associates one or more labels. To identify the category or class to which a text is associated, a set of steps is usually followed. These steps mainly concern how a text is represented, the choice of the learning algorithm to use

and how to evaluate the results obtained to ensure a good generalization of the learned model.It has two phases that can be distinguished as follows:

Learning, which involves several steps and results in a model of prediction

1-We have a set of texts labelled

2- From this corpus, we extract the k descriptors (t1; ..; tk) most relevant in the sense of the problem to be solved

3- We then have a table of "descriptors - individuals", and for each text we know the value of its descriptors and its etiquette

4- We apply a learning algorithm on this chart to obtain a prediction model.

The classification of a new dx text, which includes two steps

a. Search and then weight occurrences (t1; ...; tk) of terms in the text dx to be classified

b. Applying the model to these occurrences to predict the etiquette of this dx text.

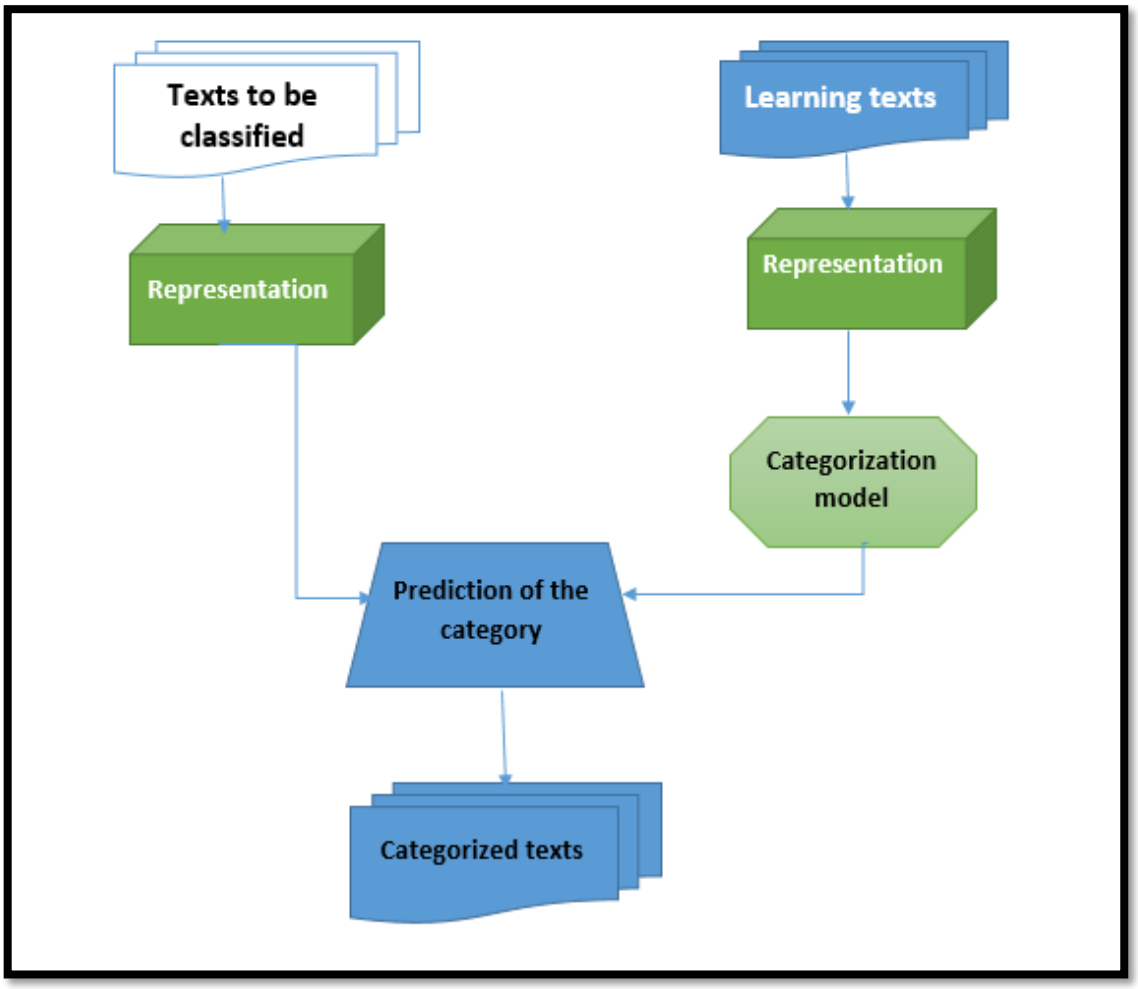


Figure 2.6: The categorization task. [22]

Note that the k most relevant descriptors (t1; ...; tk) are extracted during the first phase by analysis of the texts in the learning corpus. In the second phase, that of the classification of a new text, we simply seek the frequency of these k descriptors (t1; ...; tk) in this text to be classified. [22]

2.6.1 Representation of texts :

In order to be able to train a machine on such data, it is first of all necessary to specify an input format which is understood by the learning algorithm. For the automatic categorization of texts, the most used format is the "bag of words" vector representation.

They represent a text d in the following form :

$$d = < W_1 , W_2, \ldots\ldots W_{|T|} >$$

Where T denotes the set of terms in the lexicon and wi denotes the weighting of the term i in the text.

To calculate the weight wi the most used method is TFIDF (acronym for "term frequency inverse document frequency"). This gives more importance to the words which often appear inside the same text, and also gives less weight to the words which belong to several texts, The weight of a term t_k in a text dj is calculated so :

$$\text{tfidf}\ (t_K, d_j) = \#(t_K, d_j).\log\frac{|T\ r|}{\#(t\ _k)}$$

Or :

- $\#(t_k , d_j)$ is the frequency of T_k in d_j.
- $|T_r|$ is the number of training texts.
- $\#(t_k)$ is the number of training texts in which t_k appears at least once. [22]

2.6.2 Pretreatment :

Before vector encoding, it is necessary to carry out beforehand the encoding of a document in a space of words a transformation allowing the passage of the space of the character to a space of words. This transformation is called pretreatment.[22]

2.6.2.1 Segmentation :

The segmentation phase consists of cutting out the sequence of characters in order to group the characters forming the same word. A classic segmentation consists in cutting out the sequences of characters according to the presence or the absence of separating characters (of type "space", "tabulation" or "line break").[22]

2.6.2.2 Morphological treatment :

Morphological processing consists of processing at the level of each of the words in order to group together a set of significantly identical words. [22]

- **Deletion of stop words :**

Certain words like prepositions, conjunctions or articles, which contain no semantic information, which does not change the meaning of the words which accompany them,

could be removed from the documents. This pretreatment, allowing the size of the lexicon to be reduced. [22]

- **The stemming:** Consists in grouping under the same identifier words whose root is common. [22]
- **Lemmatization** : Lemmatization consists in grouping together words whose meaning is the same even though their roots are different (for example, “house” and “barrack”). It is a more complex task than stemming and usually involves the use of large knowledge bases. Example : the lemma of a plural word will be its singular form,that of a conjugated verb will be its infinitive form.[22]

2.6.2.3 Syntactic processing :

2.6.2.3.1 Selection of attributes : The most used criteria are :

- **Frequency**: It is simply a matter of eliminating words whose number of documents in which they appear is below a certain threshold. The underlying idea is that these words do not provide useful information for predicting the category of a text or that they do not influence the overall performance of the classifier. There is also the possibility that these terms are the result of errors, such as a misspelled word. In this case, their elimination is therefore beneficial. Operating this way to reduce the number of attributes is fairly simple and quick. One caveat to using this technique is that generally it is not used for really aggressive selection, because it is recognized that terms with low / medium frequency are relatively informative and should be kept. [22]

- **Information gain** : We measure in a way the power of discrimination of a word, the number of bits of information obtained for the prediction of the category knowing the presence or absence of 'a word. This method is often used in decision trees, to choose the attribute that best divides the set of instances into two homogeneous groups. [22]
- **Mutual information:** This way of evaluating the quality of a word in predicting the class of a document is based on the number of times a word appears in a certain category. The more a word appears in a category, the higher the mutual information of the word and the category will be judged.

The more a word will appear outside the category (and the more category will appear without the word), the lower the mutual information will be deemed to be high. You then have to average the scores of the word combined with each of the categories. The weakness of this measure is that it is much too influenced by the frequency of words. For the same conditional probability knowing the category, a rare term will be favored,because it is less likely to appear outside the category.[22]

- **The statistics of X^2 :** A well-known statistical measure, it adapts well to the selection of attributes because it assesses the lack of independence between a word and a class. It uses the same notions of word / category co-occurrence as mutual information, but an important difference is that it is subject to standardization, which makes the terms more comparable. It still loses relevance for infrequent terms. [22]
- **The strength of the term :** This is a rather different method from the others. She sets out to estimate the importance of a term based on its propensity to appear in similar documents. A first step consists in forming pairs of documents whose cosine similarity is greater than a certain threshold. The strength of a term is then calculated using the conditional probability that it will appear in the second document of a pair, knowing that it appears in the first.[22]

2.6.2.3.2 Table "individuals - variables" : Consists of transforming all the texts into a cross table "individuals - variables":

- **The individual**: Is a dj text, labeled during the learning phase, and to be classified in the prediction phase. [22]
- **The variables**: Are the descriptors (the terms) tk which are extracted from the textual data. [22]
- **The content of the array** (the elements wkj), At the intersection of the text j and the term k, represents the weight of this term k in the text j. [22]

2.7 Analysis of a NALP system :

At this level, two formulas have been carried out. A little old, in terms of morphology and syntax, and the other much more recent in terms of semantics and linguistic pragmatics. It should be noted that lexical semantics, which explain the meaning of

individual units, are often confused with propositional semantics, which studies the meaning of statements as a whole and which can be given a value of truth. [23]

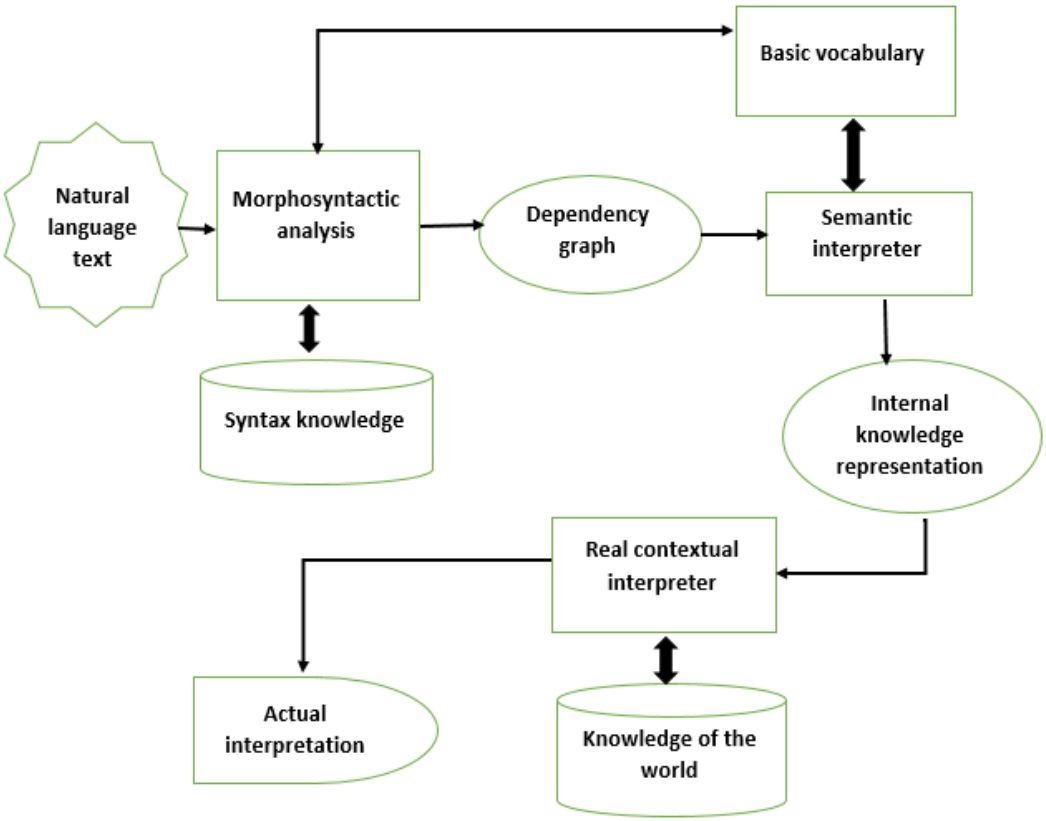


Figure 2.7: General architecture of the NALP. [23]

- ✓ **The morphological analyzer(AM) :**
  - **Morphology** :

Interprets how words are structured and what their roles are in the sentence. This analysis consists of a segmentation of the text into elementary units to which knowledge is attached in the system: once this segmentation is carried out, it is no longer the text which is manipulated, but an ordered list of units. For the processing of a digital text: we start with a chain of typographical characters, and we try to segment it so that each part corresponds to a unit classified in the system.[24]

- **Segmentation and analysis of the shape:**

The segmentation operation proceeds by accessing the different tables (clitics and suffixes) to detect the presence of proclitic, enclitic, prefixes in the form. The result is a set of five segments (proclitic, prefix, radical, suffix, enclitic). In addition, the analysis

operation accesses the dictionary of simple forms to verify the existence of the radical. If it exists, the analyzer associates all of its linguistic information (the base, root, etc.) with it.[25]

So,at the end of the AM of the request, the analyzer produces a set of information (bade, root, grammatical category, set of syntactic features) which represents the out-of-context morphological solution calculated in the model linguistics used. [26]

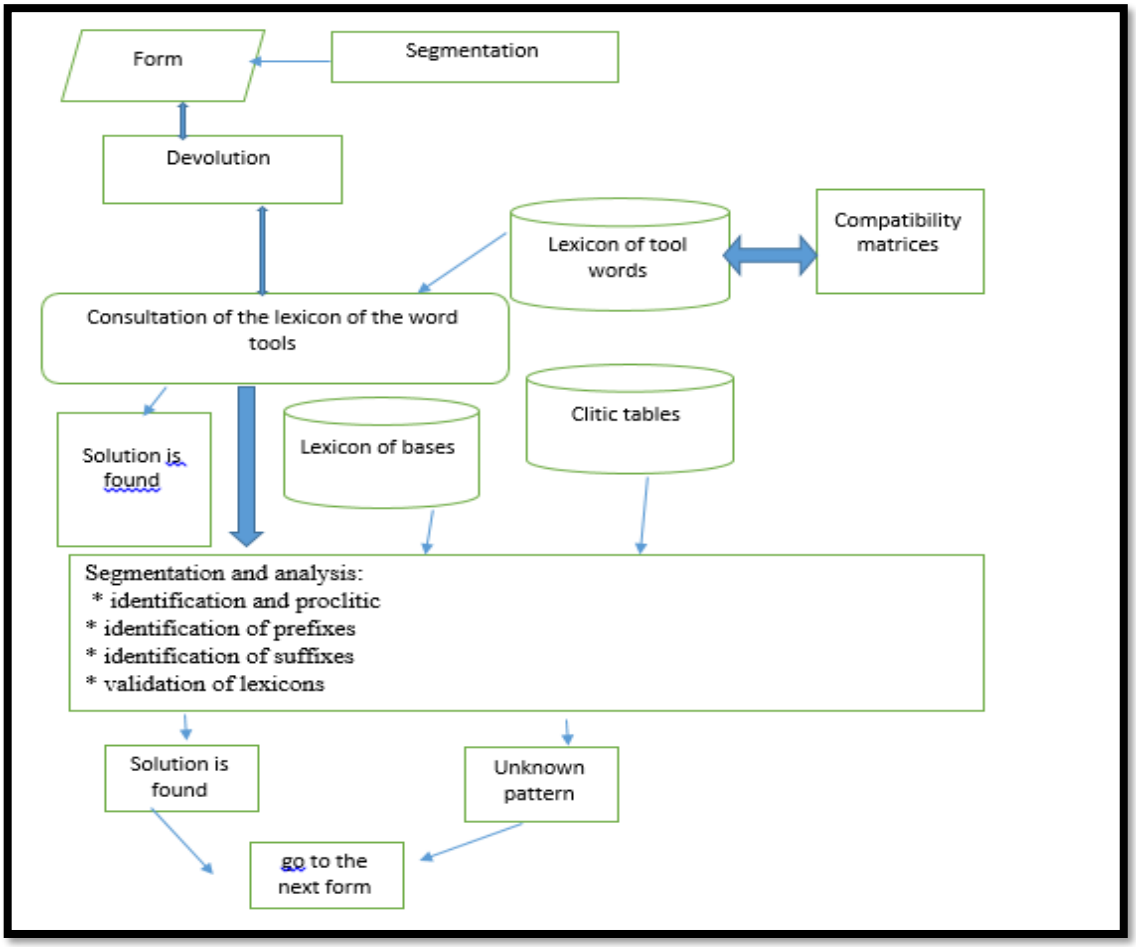


Figure 2.8: General architecture of the morphological analyzer. [24]

- ✓ **Syntactic analysis :**

It is a part of the grammar which deals with the way in which words can combine to form propositions and the sequence of propositions between them. This consists in associating, with the chain divided into units, a representation of structural groupings between these units as well as functional relationships that unite groups of units. [24]

✓ Semantic analysis :

The semantic level is still much more complex to describe and formalize than the previously stated levels. Therefore, few processing tools remain operational or at least, relate to very reduced applications where semantic analysis is limited to a perfectly narrow domain.

The sentence is the main unit of analysis that semantic processing supports in order to represent its significant part. These sentences, the meaning of which the semantic analyzer must describe, consist of a number of words identified by morphological analysis, and grouped into structures by syntactic analysis. These words and these structures constitute as many clues for the calculation of the meaning. [24]

✓ Contextual analysis :

The sentence treated out of context, that is to say isolated from its text, may not have the same meaning as in context. The semantic analysis of the isolated sentence, leads us to represent the part of the meaning of the words in this sentence, it therefore does not exhaust what can be called the complete meaning of a text, namely the existing relationships between the sentences of the text as the human apprehends it during a process of comprehension.

This is how contextual analysis comes into play, which consists in finding the "real" meaning of sentences linked to the positional and contextual conditions of use of words.

Text mining (also referred to as *text analytics*) is an artificial intelligence (AI) technology that uses natural language processing (NLP) to transform the free (unstructured) text in documents and databases into normalized, structured data suitable for analysis or to drive machine learning (ML) algorithms. [24]

2.8 Machine Learning for Natural Language Processing:

Machine learning for natural language processing and text analytics involves using machine learning algorithms and “narrow” artificial intelligence (AI) to understand the meaning of text documents. These documents can be just about anything that contains text ,the role of machine learning and AI in natural language processing (NLP) and text analytics is to improve, accelerate and automate the underlying text analytics functions and NLP features that turn this unstructured text into useable data and insights.Unlike

algorithmic programming, a machine-learning model is able to generalize and deal with novel cases. If a case resembles something the model has seen before, the model can use this prior "learning" to evaluate the case. The goal is to create a system where the model continuously improves at the task you've set it.

Machine learning for NLP and text analytics involves a set of statistical techniques for identifying parts of speech, entities, sentiment, and other aspects of text. The techniques can be expressed as a model that is then applied to other text, also known as supervised machine learning. It also could be a set of algorithms that work across large sets of data to extract meaning, which is known as unsupervised machine learning.[27]

- **Named Entity Recognition**

At their simplest, named entities are people, places, and things (products) mentioned in a text document. Unfortunately, entities can also be hashtags, emails, mailing addresses, phone numbers, and Twitter handles. In fact, just about anything can be an entity if you look at it the right way. And don't get us stated on tangential references.

At Lexalytics, they've trained supervised machine learning models on large amounts pre-tagged entities. This approach helps us to optimize for accuracy and flexibility. They've also trained NLP algorithms to recognize non-standard entities (like species of tree, or types of cancer).It's also important to note that Named Entity Recognition models rely on accurate PoS tagging from those models. [27]

- **Categorization and Classification**

Categorization means sorting content into buckets to get a quick, high-level overview of what's in the data. To train a text classification model, data scientists use pre-sorted content and gently shepherd their model until it's reached the desired level of accuracy. The result is accurate, reliable categorization of text documents that takes far less time and energy than human analysis.

ML vs NLP and Using Machine Learning on Natural Language Sentences Let's return to the sentence, "Billy hit the ball over the house." Taken separately, the three types of information would return:

- **Semantic information**: *Person – act of striking an object with another object – spherical play item – place people live*
- **Syntax information**: *Subject – action – direct object – indirect object*
- **Context information**: *This sentence is about a child playing with a ball*

These aren't very helpful by themselves. They indicate a vague idea of what the sentence is about, but full understanding requires the successful combination of all three components.

This analysis can be accomplished in a number of ways, through machine learning models or by inputting rules for a computer to follow when analyzing text. Alone, however, these methods don't work so well.[27]

2.9 Conclusion :

In this chapter we have defined the concept of social network analysis ; its historical development then cite some areas of use for this analysis. Subsequently we introduced the methods that allow the analysis of social networks, then compare them to draw the one that has the ideal qualities to analyze very large networks of size of Twitter or Facebook ... etc.

After that we presented definition of text mining and NLP, we learned how to categorize a text and And his mechanism.finally we defined Machine Learning for Natural Language Processing.

Chapter 3: Related work

3. Related work

3.1 Introduction :

In this chapter, we present works related to our research. We found 4 compilations based on the classification of religions, each according to its method.

Through our research in the references, we found works that depend on how the machine learns, including the first work" On Predicting Religion Labels in Microblogging Networks"[28] and the second work is " Detection and classification of social media‐based extremist affiliations using sentiment analysis techniques "[29], but The third work " Lexical Analysis of Religious Texts using Text Mining and Machine Learning Tools "[30] and the last work " Style of Religious Texts in 20th Century "[31] used other methods of classification that we will learn about through this chapter.

3.2 On Predicting Religion Labels in Microblogging Networks : [28]

Religious belief plays an important role in how people behave,and develop relationships with others.The religion labels of user population are obtained by conducting a large scale census study.They study the problem of predicting users' religion labels using their microblogging data.

They formulate religion label prediction as a classication task, and identify content, structure and aggregate features considering their self and social variants for representing a user.They introduce the notion of representative user to identify users who are important in the religious user community.They further dene features using representative users. They show that SVM classiers using their proposed features can accurately assign Christian and Muslim labels to a set of Twitter users with known religion labels.Online social media users have their religion attributes embedded in the content and interaction with other users. It is therefore interesting to recover these users' religion labels from their social media data, and to do so accurately

3.2.1 Objectives:

They attempt to predict users' religions using their microblogging data.This task has not been addressed so far and datasets with religion labels are not publicly available.Their

research begins with gathering a dataset covering a set of 111,767 Twitter users located in Singapore, their follow relationships, and their tweets. This Twitter dataset allows for them to explore both content and structure features relevant to users' religious beliefs.

They manually annotate the religion labels of over one thousand users who declare religion beliefs in their biographies. Using these labeled data, they are able to evaluate methods that predict user religion labels.They crawled the Twitter data generated by about 110K users with Singapore specied as their prole location in July 2012.The dataset consists of users' proles, tweets, and follow links among the users.

User religion labeling.There are five main religions in Singapore, namely, Buddhism, Taoism, Christianity, Islam, and Hinduism.When they attempted to assign religion labels to Singapore users, they found very few self-declared Buddhists, Taoists and Hindu's in their dataset.They therefore focus on labeling Christians and Muslims only. Only those who clearly state their religion were then assigned the Christian or Muslim label.They managed to label 581 Christians and 448 Muslims.This composition structure is quite different from religion composition reported in the 2010 Singapore population census which shows Christians, Muslim, Buddhists and Free-thinkers represent 18.3%, 33.3%, 14.7% and 17% respectively.

3.2.2 User Feature Representation :

They represent each user using a vector of features which in turn will be used in learning classiers. They categorize all the features by their data type as well as by the user(s) from which the features are derived from.by considering different combinations of data types and users, they obtain a comprehensive feature set. By data type, they define: (a) Content features: which are derived from textual content of tweets; (b) Structural features: which are features derived from the connectivity of target user in relation with his neighbors; (c) Aggregated features: which are features derived from summarizing the content or structural features of the target user.

The features can also categorized by the user(s) from which the features are extracted. There are: (a) Self features which are extracted from the target user's data only; and (b) Social features which are extracted from the neighborhood which includes the target user and other (possibly selected) users connected to him. Assuming that homophily

exists, users in the neighborhood are expected to have similar features that enrich the target user representation.

3.2.3 Experiments :

They evaluate the performance of their prediction method against the 1000+ ground truth labeled users.They use WEKA, a machine learning software for computing experiment metrics.Weka includes a wrapper of libSVM, an implementation of SVM by Chang and Lin .Using 10-fold cross validation, they obtain the performance of the classiers using different com+binations of features metrics.

Their performance metrics are the standard precision, recall and F1 scores for the Christian and Muslim classes. Performance of using content features only comparing self content with content generated in the neighborhood is depicted. The results show that social content from neighbors only is comparable to self content and social content from neighborhood outperforms both self content and social content from neighbors only.

3.2.4 Evaluation on all users :

They next evaluate their classification method on all the 110K+ users who have not been assigned any religion labels.These users do not come with ground truth labels.They therefore manually judged the top 50 scored users under the Christian class, and those top 50 scored users under the Muslim class. The manual judgement was conducted by (1) examining religion related tweets generated by the users, (2) checking their prole pages including biography, recent tweets, and (3) checking his followings and followers.The results show that these top scored users are indeed assigned with the correct religions,i.e., the precision of the manual check is 100%.

3.3 Detection and classification of social media-based extremist affiliations using sentiment analysis techniques : [29]

Identification and classification of extremist-related tweets is a hot issue. Extremist gangs have been involved in using social media sites like Facebook and Twitter for propagating their ideology and recruitment of individuals. They aim at proposing a terrorism-related content analysis framework with the focus on classifying tweets into extremist and non-extremist classes. Based on user-generated social media posts on

Twitter, they develop a tweet classification system using deep learning-based sentiment analysis techniques to classify the tweets as extremist or non-extremist.The experimental results are encouraging and provide a gateway for future researchers.

Opinions expressed on such sites give an important clue about the activities and behavior of online users. Detection of such extremist content is important to analyze user sentiment towards some extremist group and to discourage such associated unlawful acts.

They investigate deep learning-based sentiment analysis techniques,which have already shown promising performance across a large number of complicated problems in different domains like vision, speech and text analytics.

They propose to apply LSTM-CNN model, which works as follows: (i) CNN model is applied for feature extraction, and (ii) LSTM model receives input from the output of the CNN model and retains the sequential correlation by taking into account the previous data for capturing the global dependencies of a sentence in the document with respect to tweet classification into extremist and non- extremist.

They take the training set Tr = {t1, t2, t3,…..tn} and class tags (labels) has Extrimist_affliation ={yes,no}.

Each tweet is assigned a tag.The aim is to design a model which can learn from the training data set and can classify a new tweet as either extremist or nonextremist.

They apply IBM Watson API for tone analysis.And They experiment with multiple Machine Learning (ML) classifiers such as Random Forest, Support Vector Machine, KN Neighbors, Naïve Bayes Classifiers. The feature set for such classifiers is encoded by taskdriven embedding trained over different classifiers : CNN, LSTM, and CNN + LSTM.

3.3.1 Data collection :

They used Twitter streaming API to scrap tweets containing one or more extremismrelated keywords. Furthermore, they also investigated different Dark Web forums, such as Al-Firdaws, Montada, alokab, and Islamic Network.

The first three are Arabic forums and third one is in English. They collected over 25,000 postings by translating the non-English postings to English using Python-based Google Translate API. Each review is matched with the seed words present in the manually built extremist's vocabulary lexicon acquired from BiSAL, a bilingual sentiment lexicon for analyzing dark web forums. In this way, all postings containing one or more keywords from the manually acquired lexicon, are collected.

3.3.2 Preprocessing

They applied different preprocessing techniques, such as tokenization, stop word removal, case conversion, and special symbol removal. The tokenization yields a set of unique tokens (356,242), which assist in building a vocabulary from the train-ing set, used for encoding the text.

Tweet id	Tweet	Tweet label
1	Oh Allah, we are helpless	Non-extremist
2	Great news, ISIS fight Afghan forces to capture Helmand..	Extremist
3	Love you Baghdadi, Who is interested to know the truth about ISIS Yes.. visit our page to listen # Montada	Extremist
4	A very painful fight……operation zarbe-azab gave us a big blow #TTP	Extremist
5	What a great news, a suicide bomber destroyed the police station and killed 20 people	Extremist

Table 3.1: A partial listing of training data.

3.3.3 Training, validation and testing :

They divided the dataset into three parts : train, validate, and test. The DL model is trained with the Keras library based on TensorFlow. The hardware requirements include 4 Titan X GPUson, a 128 GB memory with Intel Core i7 node .Training data Training data is used to train the model.80% of the data is used training and it may vary as per requirements of the experiment. The training data includes both the input and the expected output. It includes both the input and the corresponding expected output.10% validation set is used to avoid performance error by applying parameter tuning. For this purpose, they applied automatic verification of dataset, which provides an unbiased evaluation of the model and minimize the overfitting.

3.3.4 Network model :

The proposed method implements and evaluates the performance of long short term memory with Convolutional Neural Network (CNN) model to identify tweets/reviews containing content with extremist clues.They train the neural classifier, for the classification of extremist affiliation content.The working flow of the network is comprised of following steps :

(i) word embedding, in which each word in a sentence is assigned a unique index to form a fixed-length vector, (ii) dropout layer is used to avoid overfitting, (iii)LSTM layer is incorporated to capture long-distance dependency across tweets/reviews (iii) feature extraction is performed using a convolution operation, (iv) pooling layer aims at minimizing the dimension of feature map by aggregating the information,(v) Flatten layer converts the pooled feature map into a column vector, and (vi) at the output layer, softmax function is used for the classification.

3.3.5 Analyzing sentiments of users w.r.t emotional affiliation with extremists :

This module deals with emotion classification of the user's sentiments showing affiliation with the extremists' postings.Each of the input text is tagged with an emotion category using the Python-based tone analyzer API.It returns an emotion set: {anger, sadness, fear, joy, confident, analytical and tentative}.In this module, an input text is analyzed and the corresponding emotion class is identified.

3.3.6 Experimental setup

3.3.6.1 Results and discussion

How to recognize and classify tweets as extremist vs non-extremist, by applying deep learning-based sentiment analysis techniques ?

To answer this research question, they applied deep-learning-based sentiment analysis technique, namely LSTM + CNN.

They conducted experiments on 1-Layer CNN, 1-Layer LSTM. Models and it is obvious that the proposed LSTM + CNN model achieves best results.At the next level, the CNN layer captures additional features (n-gram) from the richer collection, yielding better and improved performance results.

Table 3.2 shows results obtained on account of applying different DL models and it is obvious that the proposed LSTM + CNN model achieves best results.

Models	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
1-layer LSTM	85.07	87	85	85
1-layer CNN	83.09	84.4	82	82
LSTM + CNN	92.66	88.32	89.47	90.71

Table 3.2 : Experimental results of different DL-based SA models.

Table 3.3 presents the parameter setting for the proposed model (LSTM+ CNN). The experiment is conducted using different parameters, listed as follows: the parameter, namely 'number of filter', receives a value varying from 2 to 16, while the values of other parameters, such as kernel size, padding, pooling size, optimizer, batch size, epochs, and units, are fixed.

LSTM + CNN model layers	Parameters with values
Convolutional layer	Number of filters = 2, 4, 6, 8, 10, 12, 14, 16 Kernel size = 2 × 2 Padding = same
Pooling layer	Pooling size = 2 × 2
LSTM layer	Units = 100
Additional	Vocabulary size = 2000 Input vector size = 50 Embedding dimension = 128 Batch size = 8 Number of epochs = 7

Table 3.3 : Parameter setting regarding proposed LSTM+CNN model .

3.3.6.2 Parameter settings :

Table 3.3 presents the parameter setting for the proposed model (LSTM + CNN). The experiment is conducted using different parameters, listed as follows:

The parameter, namely 'number of filter', receives a value varying from 2 to 16, while the values of other parameters, such as kernel size, padding, pooling size, optimizer, batch size, epochs, and units, are fixed.

3.3.6.3 Deep learning methods with variants

Different variants of DL techniques (CNN +Random Embedding, LSTM+ Random Embedding, FastText + Random Embedding, and GRU + Random Embedding), are used for tweet classification as extremist and non-extremist. The results are reported in

Table 3.4. The proposed LSTM + CNN performs better than the other DL methods. [28]

Study	Technique	Results		
		P	R	F
Part-A Baselines	N-gram + SVM [12]	73	78	79
	Bag of words +SVM [13]	73	74	73
	TF-IDF +RF	78	84	90
	BOW + KNN [8]	76	75	76
Part-B Deep learning methods with variants	CNN + word embedding	84	84	84
	LSTM + word embedding	86	86	86
	FastText + word embedding	87	87	87
	GRU + word embedding	85	85	85
Part-C Proposed	LSTM + CNN	0.90	0.88	0.88

Table 3.4: Comparison of proposed word with baseline methods.

3.3.6.4 The method

For the extremist classification task, they experimented different DL-based SA models, namely CNN, LSTM, FastText and GRU, initialized with random word embeddings. The proposed LSTM + CNN model for extremist classification outperforms baseline methods and it also performs better than the other DL models.They conducted experiments using supervised,lexicon-based, benchmark proposed technique for extremist Classification on Twitter.

3.3.6.5 Supervised techniques :

In Table 3.5, the results of the proposed method (LSTM + CNN), are compared with different state-of-the-art techniques [7,8] based on machine classifiers, namely, KNNeighbors, Naïve Bayes Classifier.Furthermore,they also applied other machine and deep learning classifiers, namely Random Forest, Support Vector Machine, LSTM, and CNN.The performance evaluation results of various machine learning classifiers are presented in terms of accuracy, precision, recall, and f-measure. The KNN exhibited the lowest performance result (72% accuracy).

Study	Techniques	Features	P (%)	R (%)	F (%)	A (%)
Wei et al. [8]	Machine learning (KNN)	Classical features (tf, idf, tfxidf) BOW	73	71	71	74
Azizan and Aziz [7]	Machine learning (NB)	Classical features (tf, idf, tfxidf)	70	68	69	72
Proposed	Deep learning (LSTM + CNN)	Word embedding	90	86	84	92
Chalothorn and Ellman [17]	Lexicon-based (Senti WordNet)	Sentiment scores and polarity classes	69	68	69	73
Other Models	Deep learning (LSTM)	Word embedding	85	84	83	85
	Deep learning (CNN)	Word embedding	88	84	83	88
	Machine learning (SVM)	Tf xidf	79	78	79	79
	Machine learning (RF)	Tf xidf	83	81	82	84

Table 3.5: Comparison with state-of-art technique

The proposed technique (LSTM+ CNN) produced the best results (Table3.5), when compared with the other comparing methods.

At the end presents a sentiment-based extremist classification system based on users' postings made on Twitter. The proposed work operates in three modules: (i) users' tweet collection, (ii) preprocessing, and (iii) classification with respect to extremist and nonextremist classes using LSTM + CNN model and other ML and DL classifiers.

The experimental results show that the proposed system outperformed the comparing methods in terms of better precision, recall, f-measure, and accuracy.

3.4 Lexical Analysis of Religious Texts using Text Mining and Machine Learning Tools : [30]

They presents a text mining approach to compare and to explore the similarities and the differences between various religious texts using POS Tagging and Term Document Matrix. Automated text mining and machine learning tools have been used for lexical analysis of the ten world famous religious texts :The Holy Bible, the Dhammapada, the Tao Te Ching, the Bhagwad Gita, the Guru Granth Sahib, the Agama, the Quran, the Rig Veda, the Sarbachan and the Torah.The extracted nouns categories were used as features to explore some interesting relationships between these religions and ideas that have emerged in different religions from different geographic regions.

3.4.1 Collection Of Religious Texts

The English translations of religious texts were downloaded from the Internet in the pdf format. The books analyzed included the Holy Bible, the Dhammapada, the Tao Te

Ching, the Bhagwad Gita, the Guru Granth Sahib, the Agama, the Quran, the Rig Veda, the Sarbachan and the Torah.

3.4.2 Parts Of Speech (Pos) Tagging Of Text

After selecting the books for the lexical analysis, Part-ofspeech tagging (POS tagging) of the words, constituting the religious texts was done using the "openNLP" package in R. The Part-of-speech tagging is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context.The POS tags used by the "openNLP" package are the Penn English Treebank POS tags.

POS tagging was followed by extraction of lexical tokens or words from the texts was done using R. Lexical tokens are the words which primarily convey information, which include nouns, verbs, adjectives and adverbs.

3.4.3 Calculation of lexical richness of religious texts :

The most popular measures used in the measurement of Lexical Richness are Lexical Density(LD) and Lexical Variation(LV).Lexical Density is defined as the percentage of lexical words in the text ,i.e. nouns ,verbs ,adjectives ,adverbs:

$$\text{LD} = \frac{Number\ of\ lexical\ tokens}{Total\ number\ of\ token} * 100$$

lexical Variation(LV) is the type/token ratio, i.e.The ratio in percent between the different words in the text and the total number of running words:

$$\text{LD} = \frac{Number\ of\ type}{Total\ number\ of\ tokens} * 100$$

The Lexical Density(LD) and Lexical Variation(LV) of the ten religious texts was calculated as shown in Table 6. The observations made are as follows :

- The Agama had the highest Lexical Density followed by the Rig Veda and the Tao Te Ching. The Torah had the least Lexical Density as shown in Figure 3.27.
- The Bhagwad Gita had the highest Lexical Variation followed by the Agama and the Tao Te Ching.The Guru Granth Sahib had the least Lexical Variation .

Categories	Agama	Bhagwad Gita	Holy Bible	Dhammapada	Guru Granth Sahib	Quran	Rig Veda	Sarbachan	Tao Te Ching	Torah
Family Relationships	Mother,son,child,father,wife	Son,mother,father ,children, wife	Son,father,mother, husband, children	Mother,father,son ,wife,children	Husband, mother,child,father,wife,daughter	Mother,son,father, children, wive,brethren	Mother,son,father, child,husband,brother,sister	Mother,husband,father,son ,brother, wife,child	Mother,son,children	Mother,father,children,son, wife,child
External Body Parts	Head,eye ,feet,hair, hnd,mouth,knee	Eye,mouth,feet,arm,palm, teeth,ear, hair	Hand,feet,eye,head,face,mouth,tongue	Tongue,hair,arm,hand,feet, head,eye, mouth,ear	Feet,tongue,forehead,eye,mouth,palm	Hand,eye ,face,feet ,ear,head ,palm,mouth,	Arm,tongue,eye, mouth,foot,hair,teeth,ear	Feet,eye, head,neck,head,er ,arm,tongue,mouth	Arm,eye, ear,mouth,foot,hand	Teeth,palm,head,shoulder,ear,face,eye.feet
Internal Body Parts	Mind,heart,limb,womb,bone,blood	Flesh,mind,heart, womb,blood	Heart,mind,womb ,bone,blood,breast	Heart,mind,bone, blood	Mind,heart,womb, bone,breast,blood	Heart,womb,bone, blood,breast	Heart,mind,limb, womb,bone,blood, breast	Mind,heart,breast, womb,bone	Mind,heart,bone	Flesh,blood,womb ,breast,bone,heart
Emotions	Anger,pain,pleasure,love,hatred,freedom,joy, sorrow	Pleasure, pain,fear, love,anger,happiness ,sorrow,freedom,hope,honor,danger, ego,pride ,hate	Pain,fear, joy,hope, love,honor,hate,danger,anger,terror ,pride,wrath	Pain,pleasure,happiness,fear,love,hatred,anger,sorrow ,freedom, joy,danger,honor	Love,pain,egotism,fear,hope,anger, danger	Pleasure, hate,fear, wrath,danger	Joy,delight,love,pleasure,freedom,anger,danger,fear,hope,sorrow,terror, pain	Plesaure, pain,fear, anger,honor,sorrow,laugh, delight,distress	Honor,fear,distress,sorrow, danger,love	Terror,dread,pain, sorrow,joy,love,anger,danger
Virtues	Kind,truth,faith,wisdom,strength,care,knowledge	Sacrifice, wisdom, art,grace, kind,,righteous,wit	Mercy,courage,wisdom,strength,truth,grace,faith,wit,care	Truth,wisdom,strength,kind,care,trust	Grace,wisdom,grace,mercy ,kind,faith,courage	Wisdom, faith,sacrifice,strength	Art,wisdom,kind, grace,care,wrath	Mercy,art,devotion,grace,faith,strength,wisdom,courage	Righteousnes,sacrifice,trust ,wit,wisdom,care, wit	Grace,wrath,courge,kind,care,wisdom
Date/Time	Day,dark ,night,hour,month ,dawn,year,today	Dark,night,day,month,dawn,today,year	Day,night,morning,dark,hour,year,month,spring,today	Night,day,dark,month,hour,year,today	Night,day,dark,hour,morning,dawn	Day,night,hour,morning,dark,dawn, month	Dawn,day,morning,night,dark,today	Night,day,hour,today,morning,dawn ,month	morning	Month,morrow,today,day, night
Plants	Plant,fruit,tree,flower,root	Wood,lotus,tree,root,seed,plant,herb,forest,grass	Tree,grass,forest,herb,plant, branch,seed	Forest,flower,lotus,tree,grass,sandalwood,plant,wood,root,thorn ,seed	Lotus,forest,plant, flower,sandalwood,wood,herb	Tree,plant,seed,thorn,herb,flower,root	Grass,plant,wood,forest,tree ,herb,seed,lotus	Lotus,flower,seed, tree,plant ,wood,forest,sandalwood	Plant,tree ,thorn,root	Flower,grass,plant ,forest,root,tree
God	Lord,guru,God,Goddess,Sri,Mahavira	Lord,God,Vishnu ,Shiva,Rudra,Yama,Indra, Varuna,Soma,Kali ,Krishna, Arjuna,bhagwan	Lord,God,Jesus,Christ,goddess	God,Lord	Lord,Guru,Nanak, God,Creator	Allah,Lord,God	Indra,god,soma,lord,varuna,indu,surya,rudra,brahaspati,goddess,visvedeva,saraswati,yama,kali,vishnu	Guru,Radhasoami ,Lord,God,Soami	Lord	Lord,God

Eatables	milk	Fruit,water,food,butter,milk,salt,nectar	Water,fruit,cake,bread,salt,sweet,corn,wheat,vine,milk,food,meat	Water,fruit,milk,honey	Water,fruit,milk,sweet,grain,butter,corn,honey,salt,bread,cake,wheat,grape,meat	Water,food,grain,vine,milk,meat,honey,grape,bread	Water,juice,milk,butter,grain,cake,corn,fruit,barley,honey,meat	Water,milk,butter,fruit,grain,sweet,cake,salt	Water,food	Milk,cake,honey,vine,salt,barley,wheat,flour
Elements of Nature	Moon,sun,air,light,rain,fire,star,river,sky,wind,ocan,light,sea	Fire,light,sun,ocean,moon,wind,air,river,rain,lake,sky,flame,star	Sky,ocean,air,wind,sun,river,light,sea,rain,moon	Moon,wind,fire,air,rain,light,sun,sky,star,sea	Light,ocean,fire,sun,rain,moon,wind,river,air,sea	Fire,sky,light,river,moon,wind,sea,ocean,	Sun,moon,light,sky,river,rain,air,wind,fire,star	Sun,mon,ocean,sky,light,flame,rain,fire,star,river,sea	River,ocean,sky,light,air,wind	Sun,rain,moon,star,flame,sky,sea,river,ocean
Events	Birth,holi,wear,heaven,marriage	Death,battle,war	Death,battle,marriage,birth	Death,birth	Death,holi,birth,Basant,battle	birth	Holi,battle,birth,marriage	Holi,death,birth,Basant,war	Death,battle,war	,marriage,birth
Precious Stone/ Metal	gold	Gold,iron,pearl,ornament	Gold,silver,iron,brass,pearl,diamond,ornaments	Gold,silver,iron,diamond,jewel	Gold,silver,iron,brass,diamond,jewel,ornament	Gold,silver,brass,pearl,ornament	Gold,iron,brass,pearl,jewel,ornament	Gold,silver,iron,diamond,pearl,jewel,ornament	-	Gold,jewel,brass,ornament
Animals	Elephant,cow,horse,bear,deer,snake,fish,bird,cattle,lion	Cow,horse,elephant,serpent,fish,dog,bee,calf,bird,lion,cattle	Lamb,horse,cattle,bird,lion,bear,dragon,calf,herd,sheep,dog,serpent,worm,camel,bull,hawk	Elephants,horse,bird,swan,fish,cow,bee,calf	Herd,bear,bull,fish,elephant,bird,serpent,horse,dear,swan,lion,cow,worm,dog,hawk,calf,sheep,bee,camel,dragon	Bird,bee,goat,bear,fish,dog,calf,cattle,horse,elephant,cow,serpent,herd	Horse,cow,bull,cattle,bird,bear,dragon,deer,herd,calf,swan,serpent,dog,elephant,lion,bee,sheep,fish,oxen	Bird,fish,snake,serpent,swan,bee,elephant,dog,lion,deer,bear,cow,calf	Horse,dog,fish,bird	Herd,bird,sheep,serpent,camel,worm,dog,hawk,lion,snake,bee,cow,bull,shekel,lamb,goat

Table3.6: Important noun categories with instances across all religions book

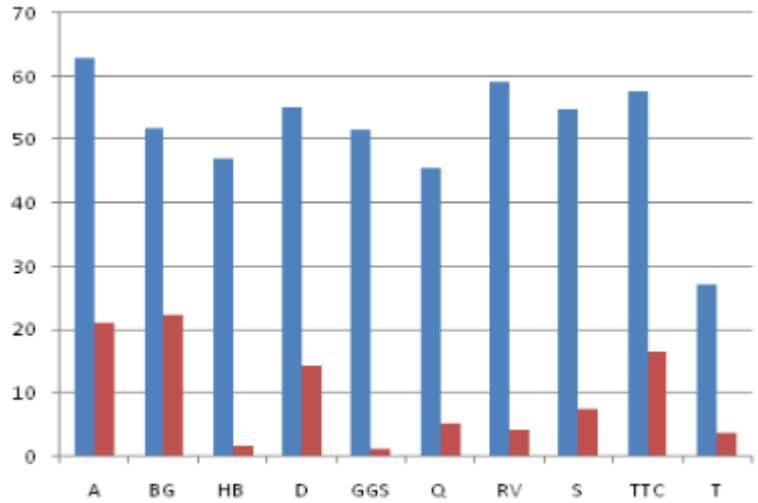


Figure 3.1: Comparison of religious texts based on lexical density and lexical variation (blue bar= lexical density, red bar =lexical variation).[30]

3.4.4 Categorisation Of Common Nouns :

In order to find the important noun categories of all the religious texts, term document matrix has been created using the “tm” package in R.The term document matrix describes the frequency count of words among all resumes. In the term document matrix, each row represents one religious text and each column represents a word (noun) and each entry represents the frequency count of a particular word in that particular resume. Some pre processing steps have been used to create the term document matrix.

Finally the results of the lexical analysis of religious texts clearly show that the religious texts originating from the same geographic region had maximum number of shared noun categories.This indicates that geographic regions had an impact on these religious texts. Some religious texts had references to the god of other religions indicating that those religious texts compared themselves with other religions of the same geographic region.

3.5 Style of Religious Texts in 20th Century : [31]

They present the results of the investigation of diachronic stylistic changes in 20th century religious texts in two major English language varieties – British and American. They examined a total of 146 stylistic features, divided into three main feature sets: (average sentence length, Automated readability index, lexical density and lexical richness), part-of-speech frequencies and stop-words frequencies. All features were extracted from the raw text version of the corpora, using the state of-the-art NLP tools and techniques.The results reported significant changes of various stylistic features belonging to all three aforementioned groups in the case of British English (1961 1991) and various features from the second and third group in the case of American English (1961–1992). The comparison of diachronic changes between British and American English pointed out very different trends of stylistic changes in these two language varieties. Finally, the applied machine learning classification algorithms indicated the stop-words frequencies as the most important stylistic features for diachronic classification of religious texts in British English and made no preferences between the second and third group of features in diachronic classification in American English.

3.5.1 Introduction :

They made a more in depth analysis of stylistic changes by using more features (a total of 146 features) and applied machine learning techniques to discover which features underwent the most drastic changes and thus might be the most relevant for a diachronic classification of this text genre.

3.5.1.1 Corpora

Their goal was to investigate diachronic stylistic changes of 20th century religious texts in British and American English and then compare the trends of reported changes between these two language varieties. Therefore, they used the relevant part (genre D – Religion in Table 3.7) of the only publicly available corpora which fulfills both conditions of being diachronic and comparable in these two English language varieties – the ‘Brown family’ of corpora.

The British part of the corpora consists of the following three corpora :

- The Lancaster1931 Corpus (BLOB),
- The Lancaster-Oslo/Bergen Corpus (LOB)
- The Freiburg-LOB Corpus of British English (FLOB).

These corpora contain texts published in 1931+-3, 1961 and 1991, respectively. The American part of the ‘Brown family’ of corpora consists of two corpora :

- The Freiburg-LOB Corpus of British English (FLOB).These corpora contain texts published in 19313, 1961 and 1991, respectively. The American part of the ‘Brown family’ of corpora consists of two corpora:
- The Brown University corpus of written American English (Brown)
- The Freiburg - Brown Corpus of American English (Frown).

Category	Code	Genre
PRESS	A	Press: Reportage
	B	Press: Editorial
	C	Press: Review
PROSE	D	Religion
	E	Skills, Trades and Hobbies
	F	Popular Lore
	G	Belles Lettres, Biographies, Essays
	H	Miscellaneous
LEARNED	J	Science
FICTION	K	General Fiction
	L	Mystery and Detective Fiction
	M	Science Fiction
	N	Adventure and Western
	P	Romance and Love Story
	R	Humour

Table 3.7: Structure of the corpora

The other four Prose genres (E–F). Although this genre contains only 17 texts of approximately 2,000 words each, the texts were chosen in the way that they cover different styles and authors of religious texts.

Although the size of the corpora used in this study (approx. 34,000 words in each corpus) is small by the present standards of corpusbased research, it is still the only existing diachronic comparable corpora of religious texts. Therefore, they find the results presented in this paper relevant though they suggest that they should be considered only as preliminary results until a bigger comparable corpora of religious texts become available.

3.5.1.2 Features :

They focused on genre D (Religion) of the four publicly available parts of the 'Brown family' of corpora and investigated diachronic stylistic changes in British (using the LOB and FLOB corpora) and American (using the Brown and Frown corpora) English. As the LOB and FLOB corpora cover the time span from 1961 to 1991, and the Brown and Frown corpora from 1961 to 1992, they were also able to compare these diachronic changes between the two language varieties in the same period 1961–1991/2. In both

cases, they used three sets of stylistic features.The first set contains features previously used by ˇStajner and Mitkov (2011) :

- Average sentence length (ASL).
- Automated Readability Index (ARI)
- Lexical density (LD).
- Lexical richness (LR).

3.5.2 Methodology

As some of the corpora were not publicly available in their tagged versions, they decided to use the raw text version of all corpora and parse it with the state-of-the-art Connexor's Machinese syntax parser, following the methodology for feature extraction proposed by ˇStajner and Mitkov (2011).

They agree that this approach allows theim to have a fairer comparison of the results among different corpora and to achieve a more consistent, highly accurate sentence splitting, tokenisation, lemmatisation and part-of-speech tagging the focus in this study will be on the POS tagging.

For each experiment, they calculated the statistical significance of the mean differences between the two corresponding groups of texts for each of the features.

Statistical significance tests are divided into two main groups: parametric and non-parametric.

In the cases where both samples follow the normal distribution, it is recommended to use parametric tests as they have greater power than the non parametric ones. Therefore, they first applied the Shapiro-Wilk's W test (Garson, 2012a) offered by SPSS EXAMINE module in order to examine in which cases/genres the features were normally distributed.

This test is a standard test for normality, recommended for small samples. It shows the correlation between the given data and their expected normal distribution scores. If the result of the W test is 1, it means that the distribution of the data is perfectly normal. Significantly lower values of W (0.05) indicate that the assumption of normality is not met.

Following the discussion in (Garson, 2012b), in both experiments they used the following strategy : if the two data sets they wanted to compare were both normally distributed they used the t-test for the comparison of their means.

After that, they applied several machine learning classification algorithms in Weka ian H. Witten, 2005) in order to see which group of the features would be the most relevant for diachronic classification of religious texts. In order to do so, they first applied two well-known classification algorithms : Support Vector Machines (Platt, 1998; Keerthi et al., 2001) and Naive Bayes (John and Langley,1995) to classify the texts according to the year of publication (1961 or 1991/2),using all features which reported a statistically significant change in that period. The SVM (SMO in Weka) classifier was used with two different settings. The first version used previously normalised features and the second – previously standardised features.

Furthermore, they tried the same classification using all possible combinations of the three sets of features: only the first set (1), only the second set (2), only the third set (3), the first and second set together (1+2), the second and third set (2+3), the first and third set (1+3). Then they compared these classification performances with the ones obtained by using all three sets of features together (1+2+3) in order to examine which features are the most relevant for the diachronic classification of this text genre.

3.5.3 Results and Discussion :

The results of the investigation of diachronic stylistic changes in religious texts are given separately for British and American English. The results of the machine learning classification algorithms for both language varieties and the discussion about the most relevant feature set for diachronic classification.

3.5.3.1 British English :

The results of diachronic changes in religious texts written in British English are given in Table3.8. The table contains information only about the features which reported a statistically significant change (sign. 0.05). The columns ‘1961’ and ‘1991’ contain the arithmetic means of the corresponding feature in 1961 and 1991, respectively. The column ‘Sign.’ contains the p-value of the applie t-test or, alternatively, the p-value of Kolmogorov-Smirnov Z test for the cases in which the feature was not normally distributed in at least one of the two years . Column ‘Change’ contains the relative

change calculated as a percentage of the feature's starting value in 1961. The sign '+' stands for an increase and the sign '' for a decrease over the observed period. In case the frequency of the feature was '0' in 1961 and different than '0' in 1991, this column contains value 'NA' (e.g. feature 'whom' in Table3.8).

Feature	1961	Sign.	Change	1991
ARI	10.332	0.010	+37.72%	14.229
LD	0.304	0.027	+8.47%	0.329
LR	0.268	0.030	+9.40%	0.293
V	14.087	0.020	−14.32%	12.070
PREP	12.277	0.016	+11.69%	13.712
A	6.843	0.009	+24.37%	8.511
ING	0.957	0.042	+31.38%	1.257
an	0.237	0.008	+52.27%	0.361
as	0.554	0.034	+32.68%	0.736
before	0.090	0.002*	−79.98%	0.018
between	0.062	0.046*	+112.07%	0.132
in	1.781	0.046*	+30.57%	2.325
no	0.230	0.027	−46.37%	0.123
these	0.191	0.017*	−45.70%	0.104
under	0.059	0.046*	−77.80%	0.013
whom	0.000	0.006*	NA	0.044
why	0.054	0.046*	−71.25%	0.015

Table 3.8: British English

The increase of ARI (Table3.8) indicates that religious texts in British English were more complex (in terms of the sentence and word length) and more difficult to understand in 1991 than in 1961.

3.5.3.2 American English

The results of the investigation of diachronic stylistic changes in religious texts written in American English are given in Table3.9 using the same notation as in the case of British English.

The first striking difference between diachronic changes reported in British English (Table3.8) and those reported in American English (Table3.9) is that in American English none of the four features of the first set (ASL, ARI, LD, LR) demonstrated a statistically significant change in the observed period (1961–1992), while in British English three of those four features did. Actually, the only feature which reported a

significant change in both languague varieties during the period 1961–1991/2 is the frequency of adjectives (A).On the basis of the results presented in Table3.9, they observed several interesting phenomena of diachronic changes in American English.The results reported a significant increase of noun and adjective frequency, and a significant decrease of pronoun frequency.

Feature	1961	Sign.	Change	1991
N	27.096	0.009	+10.72%	30.002
PRON	7.763	0.046*	−30.04%	5.431
A	8.092	0.040	+19.63%	9.680
ADV	6.050	0.020	−14.58%	5.168
all	0.298	0.017*	−36.99%	0.188
have	0.500	0.010	−36.52%	0.317
him	0.240	0.017*	−86.25%	0.033
it	0.871	0.015	−32.79%	0.586
not	0.661	0.046*	−13.37%	0.572
there	0.152	0.017*	−42.42%	0.088
until	0.052	0.017*	−60.14%	0.021
what	0.237	0.046*	−9.26%	0.215
which	0.054	0.046*	−48.63%	0.028

Table 3.9 : American English

3.5.3.3 Feature Analysis :

They used machine learning classification algorithms in order to examine which set of features would be the most important for diachronic classification of religious texts in 20th century.They tried to classify the texts according to the year of publication (1961 and 1991 in the case of British English, and 1961 and 1992 in the case of American English), using all possible combinations of these three sets of features: (1),(2), (3), (1)+(2), (1)+(3), (2)+(3), and compare them with the classification performances when all three sets of features are used (1)+(2)+(3). All experiments were conducted using the 5-fold cross-validation with 10 repetitions in Weka Experimenter.

The results of these experiments for British English are presented in Table 3.10.

The results (classification performances) of each experiment were compared with the experiment in which all features were used (row 'all' in Table 3.10), using the twotailed paired t-test at a 0.05 level of significance provided by Weka Experimenter.

Set	SMO(n)	SMO(s)	NB
all	90.19	92.52	85.48
(1)	68.81*	64.05*	67.86*
(2)	68.19*	70.14*	70.33*
(3)	87.24	92.67	88.81
(2)+(3)	91.14	93.33	86.48
(1)+(3)	87.38	89.52	88.43
(1)+(2)	74.48*	66.71*	70.33*

Table 3.10: Diachronic classification in British English.

From the results presented in Table 3.10 they can conclude that in British English, the changes in the frequencies of the stop words (third set of features) were the most important for this classification task.

In the case of American English, they needed to compare only the results of the experiment in which all features were used (row 'all' in Table 3.11) with those which used only the second (POS frequencies) or the third (stop-words) set of features, as none of the features from the first set (ASL, ARI, LD and LR) had demonstrated a significant change in the observed period 1961 1992. Theresults of these experiments are presented in Table 3.11.

Set	SMO(n)	SMO(s)	NB
all	73.67	77.86	72.90
(2)	70.57	67.67	71.10
(3)	74.67	76.24	78.33

Table 3.11: Diachronic classification in American English

The comparison of the results of diachronic classification between British and American English lead to the conclusion that the stylistic changes in religious texts were more prominent in British than in American English.

And in the conclusion they presented offered a systematic and NLP oriented approach to the investigation of style in 20th century religious texts. Stylistic features were divided into three main groups: (ASL, ARI, LD, LR), POS frequencies and stopwords frequencies.The results of the machine learning experiments pointed out the third set of features (stop-words frequency) as the most dominant/relevant set of features in the diachronic classification of religious texts in British English. These results were in concordance with the results of the statistical tests of mean differences which reported the highest relative changes exactly in this set of features.

Feature	1961	Sign.	Change	1991
N	27.096	0.009	+10.72%	30.002
PRON	7.763	0.046*	−30.04%	5.431
A	8.092	0.040	+19.63%	9.680
ADV	6.050	0.020	−14.58%	5.168
all	0.298	0.017*	−36.99%	0.188
have	0.500	0.010	−36.52%	0.317
him	0.240	0.017*	−86.25%	0.033
it	0.871	0.015	−32.79%	0.586
not	0.661	0.046*	−13.37%	0.572
there	0.152	0.017*	−42.42%	0.088
until	0.052	0.017*	−60.14%	0.021
what	0.237	0.046*	−9.26%	0.215
which	0.054	0.046*	−48.63%	0.028

Table 3.12 : American English

3.6 Compare the results :

3.6.1 Compare the results of work On Predicting Religion Labels in Microblogging Networks:

In the first work "On Predicting Religion Labels in Microblogging Networks", they study the problem of predicting users' religion labels using their microblogging data.

Their research begins with gathering a dataset covering a set of 111,767 Twitter users located in Singapore, their follow relationships, and their tweets. This Twitter dataset allow them to explore both content and structure features relevant to users' religious

beliefs. they manually annotate the religion labels of over one thousand users who declare religion beliefs in their biographies.In the first set of experiments, they evaluate the performance of their prediction method against the 1000+ ground truth labeled users.They use WEKA, a machine learning software for computing experiment metrics. Weka includes a wrapper of libSVM, an implementation of SVM by Chang and Lin. Using 10-fold cross validation, they obtain the performance of the classifers using different combinations of features metrics.their performance metrics are the standard precision, recall and F1 scores for the Christian and Muslim classes.their proposed RInDeg outperforms degree in deriving social content for the target users.They shall therefore use RIndeg in the subsequent experiments.The best F1 scores achieved for Christian and Muslim classes are 0.909 and 0.880 respectively.Table 3.13 shows the different combinations of feature categories.

	Content	Structural	Aggregated
Self	- binary terms - TF · IDF	- RInDeg - label degree for each religion r - label indeg for each religion r - label outdeg for each religion r - label Nindeg for each religion r - label Noutdeg for each religion r	- number of tweets - number/fraction of tweets containing selected terms for each religion r - $argmax_r$ indeg - $argmax_r$ outdeg - $argmax_r$ Nindeg - $argmax_r$ Noutdeg
Social	- binary terms from neighbors only - TF · IDF from neighbors only - binary terms from neighborhood - TF · IDF from neighborhood	- top k $\langle importance \rangle$ neighbors $\langle importance \rangle$ = {degree, RInDeg, indeg for each religion r }	- number/fraction of tweets from neighborhood containing selected terms for each religion r - avg/max/min of $\langle importance \rangle$ of top k $\langle importance \rangle$ neighbors $\langle importance \rangle$ = {degree, RInDeg, indeg for each religion r }

Table 3.13: Categorzation of features.

The results in Tables 3.14 show that combining both kinds of features indeed improves classification performance.

		Christian	Muslim
Self	**Content**	.792 .880 .834	.844 .775 .808
	Structure	.690 .986 .812	.960 .426 .590
	Aggregated	.847 .955 .898	.930 .777 .847
Social	**Content**	.830 .926 .876	.901 .837 .868
	Structure	.760 .981 .856	.961 .598 .737
	Aggregated	.769 .954 .852	.924 .728 .814
	Self + Social	.851 .976 **.909**	.902 .859 **.880**

Table 3.14: All features using RInDeg as user importance.

They manually judged the top 50 scored users under the Christian class, and those top 50 scored users under the Muslim class. The manual judgement was conducted by (1) examining religion related tweets generated by the users, (2) checking their profile pages including biography, recent tweets, and (3) checking his followings and

followers. The results show that these top scored users are indeed assigned with the correct religions,i.e., the precision of the manual check is 100%.While these empirical results are encouraging, they shall conduct a larger scale evaluation on this results in the future work.

3.6.2 Compare the results of Detection and classification of Social media‐based extremist affiliations using sentiment analysis techniques:

This work "Detection and classification of social media‐based extremist affiliations using sentiment analysis techniques"aims at proposing a terrorism-related content analysis framework with the focus on classifying tweets into extremist and non-extremist classes. Detection of such extremist content is important to analyze user sentiment towards some extremist group and to discourage such associated unlawful acts.

Training dataset is comprised of 12,754 tweets labeled as “extremist” and 8432 as “non-extremist”. Table 3.15 shows the detail of the used dataset.

Class label	# of Tweets	Avg. length of a Tweet	# of tokens with stop words	# of tokens without stop words
Extremist	12,754	10.47	86,971	74,354
Non-extremist	8432	8.92	64,183	52,485

Table 3.15: Dataset statistics

Training data Training data is used to train the model. In this work, 80% of the data is used training and it may vary as per requirements of the experiment. The training data includes both the input and the expected output. It includes both the input and the corresponding expected output .

Table 3.16 shows a sample list of review sentences in training data.

Tweet id	Tweet	Tweet label
1	Oh Allah, we are helpless	Non-extremist
2	Great news, ISIS fight Afghan forces to capture Helmand..	Extremist
3	Love you Baghdadi, Who is interested to know the truth about ISIS Yes.. visit our page to listen # Montada	Extremist
4	A very painful fight……operation zarbe-azab gave us a big blow #TTP	Extremist
5	What a great news, a suicide bomber destroyed the police station and killed 20 people	Extremist

Table 3.16: A partial listing of training data.

This module deals with emotion classification of the user's sentiments showing affiliation with the extremists' postings Each of the input text is tagged with an emotion category using the Python-based tone analyzer API.

It returns an emotion set: {anger, sadness... }. In this module, an input text is analyzed and the corresponding emotion class is identified. A sample set of user reviews and the detected emotion class is shown in Table 3.17.

Text id	Input text	Emotion class
2	Oh Allah, destroy US and Israel	Anger
3	Great news, ISIS fight Afghan forces to capture Helmand.	Joy
3	Love you Baghdadi, Who is interested to know the truth about ISIS Yes.. visit our page to listen # Montada	Joy
4	A very painful fight……clean up operation gave us a big blow #ISIS	Fear
5	A suicide bomber destroyed the police station and killed 20 people	Sadness
6	OMG…Hashish trade has provided us strong footing in achieving the 2014 fund raising target for the recurring expenditure of Khilafat #ISIS	Joy
7	Kafir governments do not allow us to transfer funds to mujahid accounts… instead sell enormously amount of heroin and opium to needy persons, earn in bitcoins and help us by donations…(Al-Firdaws)	Analytical
8	Baghdadi… our last hope, I simply love you #ISIS	Joy
9	#ISIS. today heard a sad news…our commander in Iraq has been shot dead in counter-attack…	Sadness

Table 3.17: Extremist-related emotion classification.

They will now conduct the results and discussion they conducted experiments on 1-Layer CNN, 1-Layer LSTM. Table 3.18 shows results obtained on account of applying different DL models and it is obvious that the proposed LSTM + CNN model achieves best results.The LSTM + CNN model attained best results, as the LSTM layer generates a new representation of the input tweet received from the embedding layer by capturing information

Models	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
1-layer LSTM	85.07	87	85	85
1-layer CNN	83.09	84.4	82	82
LSTM + CNN	92.66	88.32	89.47	90.71

Table 3.18: Experimental results of different DL-based SA models.

LSTM + CNN model layers	Parameters with values
Convolutional layer	Number of filters = 2, 4, 6, 8, 10, 12, 14, 16
	Kernel size = 2 × 2
	Padding = same
Pooling layer	Pooling size = 2 × 2
LSTM layer	Units = 100
Additional	Vocabulary size = 2000
	Input vector size = 50
	Embedding dimension = 128
	Batch size = 8
	Number of epochs = 7

Table 3.19: Parameter setting regarding proposed LSTM+CNN model.

Table 3.19 presents the parameter setting for the proposed model (LSTM + CNN). The experiment is conducted using different parameters, listed as follows:

The parameter, namely 'number of filter', receives a value varying from 2 to 16, while the values of other parameters, such as kernel size, padding, pooling size, optimizer, batch size, epochs, and units, are fixed.

Model	Training time (s)	Loss score	Accuracy (%)
LSTM + CNN1	23	1.25	62
LSTM + CNN2	28	1.17	74
LSTM + CNN3	44	1.11	80
LSTM + CNN4	24	0.98	88
LSTM + CNN5	19	0.98	88.12
LSTM + CNN6	48	0.97	90.01
LSTM + CNN7	44	0.99	90.4
LSTM + CNN8	29	0.96	92.66

Table 3.20: LSTM+CNN models training time, loss score and accuracy.

They have listed the accuracy, loss score, and training time in Table 3.20. After conducting experiments with varying parameter setting of LSTM+ CNN models, it is noted that the performance of LSTM + CNN8 model is better with lstm units = 100 (cells), pooling size = 2 × 2, and number of filters = 16 and it's achieved accuracy is 92.66%. It is noted that the accuracy of the model increases by increasing the number of filters.Following parameters are used for the LSTM + CNN model namely : (i) Max

features (10 000), (ii) embedding dimension (128), (iii) LSTM unit size (200), convolutional filter size (200), and (iv) a batch size 32 with 3 epochs which yielded best performance results as shown in (Part C of Table 3.21).

Study	Technique	Results		
		P	R	F
Part-A Baselines	N-gram + SVM [12]	73	78	79
	Bag of words +SVM [13]	73	74	73
	TF-IDF +RF	78	84	90
	BOW + KNN [8]	76	75	76
Part-B Deep learning methods with variants	CNN + word embedding	84	84	84
	LSTM + word embedding	86	86	86
	FastText + word embedding	87	87	87
	GRU + word embedding	85	85	85
Part-C Proposed	LSTM + CNN	0.90	0.88	0.88

Table 3.21: Comparison of proposed work with baseline methods.

The proposed technique (LSTM + CNN) produced the best results as it is showed in Table 3.22, when compared with the other comparing methods.

Study	Techniques	Features	P (%)	R (%)	F (%)	A (%)
Wei et al. [8]	Machine learning (KNN)	Classical features (tf, idf, tfxidf) BOW	73	71	71	74
Azizan and Aziz [7]	Machine learning (NB)	Classical features (tf, idf, tfxidf)	70	68	69	72
Proposed	Deep learning (LSTM + CNN)	Word embedding	90	86	84	92
Chalothorn and Ellman [17]	Lexicon-based (Senti WordNet)	Sentiment scores and polarity classes	69	68	69	73
Other Models	Deep learning (LSTM)	Word embedding	85	84	83	85
	Deep learning (CNN)	Word embedding	88	84	83	88
	Machine learning (SVM)	Tf xidf	79	78	79	79
	Machine learning (RF)	Tf xidf	83	81	82	84

Table 3.22: Comparison with state-of-art techniques.

How to perform the sentiment classification of user reviews w.r.t emotional affiliations of Extremists on Twitter and Deep Web ? To find an answer for this question, they conducted experiments using tone analyzer API for Emotion Classification of User reviews w.r.t Extremist's Affiliation on Dark Web.

Results reported in Table 3.23 show that the proposed module for emotion classification outperforms the comparing supervised machine learning classifiers in terms of improved accuracy, precision, recall, and F-measure.

Technique/classifier	Accuracy (%)	Precision (%)	Recall (%)	F1 measure (%)
Random forest (RF)	0.82	0.84	0.83	0.83
Support vector machine (SVM)	0.79	0.79	0.78	0.79
K-nearest neighbour (KNN)	0.72	0.72	0.72	0.71
Naïve Bayesian (NB)	0.71	0.70	0.70	0.69
Classification of users sentiments w.r.t emotional affiliation with extremists (proposed)	0.90	0.86	0.84	0.90

Table 3.23: Comparative results of sentiments of users w.r.t emotional affiliation with extremists.

3.6.3 Compare the results of Lexical Analysis of Religious Texts using Text Mining and Machine Learning Tools

In the third research, titled "Lexical Analysis of Religious Texts using Text Mining and Machine Learning Tools" This work presents a text mining approach to compare and to explore the similarities and the differences between various religious texts using POS Tagging and Term Document Matrix. Automated text mining and machine learning tools have been used for lexical analysis of the ten world famous religious texts : the Holy Bible, the Dhammapada, the Tao TeChing, the Bhagwad Gita, the Guru Granth Sahib, the Agama, the Quran, the Rig Veda, the Sarbachan and the Torah. The extracted nouns categories were used as features to explore some interesting relationships between these religions and ideas that have emerged in different religions from different geographic regions.

After selecting the books for the lexical analysis, Part-ofspeech tagging (POS tagging) of the words, constituting the religious texts was done using the “openNLP” package in R. POS tagging was followed by extraction of lexical tokens or words from the texts was done using R.

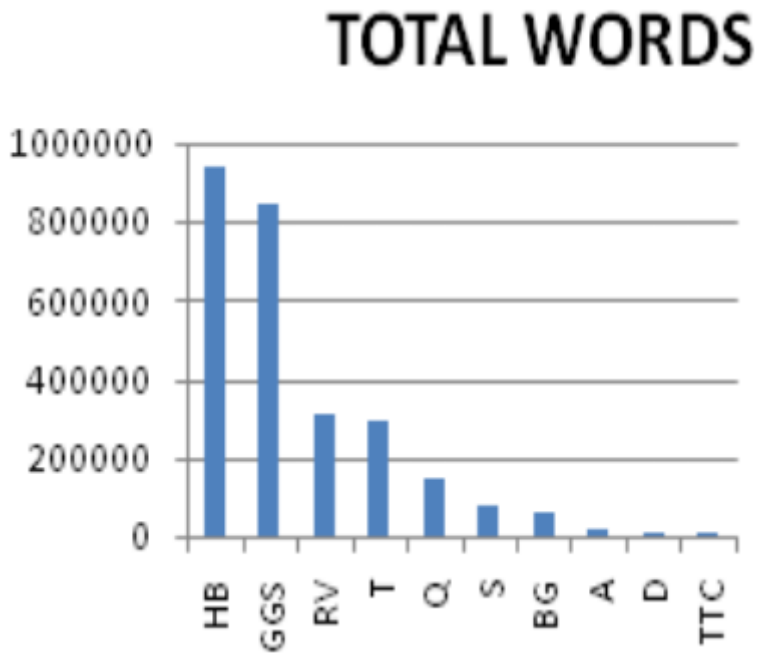


Figure 3.2: Comparison of religious texts based on total number of words.[21]

After that they will Calculation of lexical richness of religious texts,the most popular measures used in the measurement of Lexical Richness are Lexical Density(LD) and Lexical Variation(LV) Lexical Density is defined as the percentage of lexical words in the text ,i.e. nouns ,verbs ,adjectives ,adverbs

In order to find the important noun categories of all the religious texts, term document matrix has been created using the “tm” package in R. The term document matrix describes the frequency count of words among all resumes.

Some pre processing steps have been used to create the term document matrix. All the punctuation marks have been removed have been removed since these are of no significant use , all the Upper case letters have been converted into lower case in order to maintain the similarity and words are stemmed so that any differences in the tenses or the usage of the words would not be able to affect the term document matrix.The extracted nouns were categorized into common noun categories with most common.

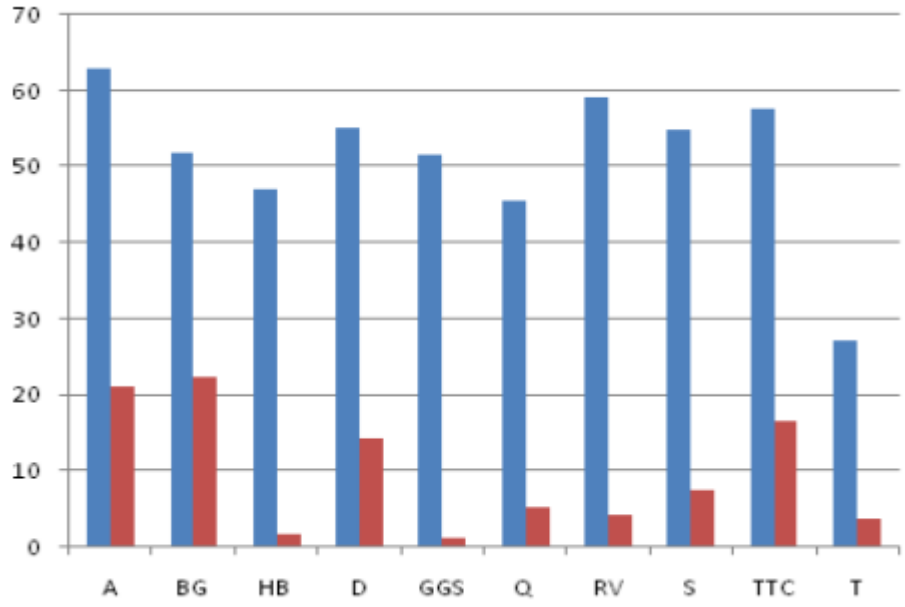


Figure 3.3: Comparision of religious texts based on lexical density and lexical variation [21]

The results of the Lexical Analysis of Religious Texts clearly show that the religious texts originating from the same Geographic region had maximum number of shared Noun categories. This indicates that geographic regions had an impact on these religious texts. Some religious texts had references to the God of other religions indicating that those religious texts compared themselves with other religions of the same geographic region.

3.6.4 Compare the results of Style of Religious Texts in 20th Century

And in the last work, titled "Style of Religious Texts in 20th Century" They conducted two sets experiments :

- Diachronic changes of style in British English .
- Diachronic changes of style in American Englis

For each experiment they calculated the statistical significance of the mean differences between the two corresponding groups of texts for each of the features.

Statistical significance tests are divided into two main groups : parametric and non-parametric.

Therefore, they first applied the Shapiro-Wilk's W test (Garson, 2012a) offered by SPSS EXAMINE module in order to examine in which cases/genres the features were normally distributed.

This test is a standard test for normality, recommended for small samples. It shows the correlation between the given data and their expected normal distribution scores. If the result of the W test is 1, it means that the distribution of the data is perfectly normal. Significantly lower values of W (0.05) indicate that the assumption of normality is not met.

Following the discussion in (Garson, 2012b), in both experiments they used the following strategy : if the two data sets they wanted to compare were both normally distributed they used the t-test for the comparison of their means; if at least one of the two data sets was not normally distributed, they used the Kolmogorov-Smirnov Z test for calculating the statistical significance of the differences between their means.

the results of the investigation of diachronic stylistic changes in religious texts are given separately for British and American English in the following two subsections.

❖ British English:

The results of diachronic changes in religious texts written in British English are given in Table3.24. The table contains information only about the features which reported a statistically significant change (sign. 0.05).

The results also demonstrated changes in the frequency of certain word types during the observed period. Verbs were used less in 1991 than in 1961, while the prepositions, adjectives and present participles were more frequent in religious texts written in 1991 than in those written in 1961 (Table3.24).

Feature	1961	Sign.	Change	1991
ARI	10.332	0.010	+37.72%	14.229
LD	0.304	0.027	+8.47%	0.329
LR	0.268	0.030	+9.40%	0.293
V	14.087	0.020	−14.32%	12.070
PREP	12.277	0.016	+11.69%	13.712
A	6.843	0.009	+24.37%	8.511
ING	0.957	0.042	+31.38%	1.257
an	0.237	0.008	+52.27%	0.361
as	0.554	0.034	+32.68%	0.736
before	0.090	0.002*	−79.98%	0.018
between	0.062	0.046*	+112.07%	0.132
in	1.781	0.046*	+30.57%	2.325
no	0.230	0.027	−46.37%	0.123
these	0.191	0.017*	−45.70%	0.104
under	0.059	0.046*	−77.80%	0.013
whom	0.000	0.006*	NA	0.044
why	0.054	0.046*	−71.25%	0.015

Table3.24 : British English.

❖ American English :

The first striking difference between diachronic changes reported in British English (Table3.24) and those reported in American English (Table3.25) is that in American English none of the four features of the first set (ASL, ARI, LD, LR) demonstrated a statistically significant change in the observed period (1961–1992).

On the basis of the results presented in Table3.25,they observed several interesting phenomena of diachronic changes in American English. The results reported a significant increase of noun and adjective frequency, and a significant decrease of pronoun frequency.

Feature	1961	Sign.	Change	1991
N	27.096	0.009	+10.72%	30.002
PRON	7.763	0.046*	−30.04%	5.431
A	8.092	0.040	+19.63%	9.680
ADV	6.050	0.020	−14.58%	5.168
all	0.298	0.017*	−36.99%	0.188
have	0.500	0.010	−36.52%	0.317
him	0.240	0.017*	−86.25%	0.033
it	0.871	0.015	−32.79%	0.586
not	0.661	0.046*	−13.37%	0.572
there	0.152	0.017*	−42.42%	0.088
until	0.052	0.017*	−60.14%	0.021
what	0.237	0.046*	−9.26%	0.215
which	0.054	0.046*	−48.63%	0.028

Table3.25: American English

They used machine learning classification algorithms in order to examine which set of features would be the most important for diachronic classification of religious texts in 20th century.

They tried to classify the texts according to the year of publication (1961 and 1991 in the case of British English, and 1961 and 1992 in the case of American English), using all possible combinations of these three sets of features: (1), (2), (3), (1)+(2), (1)+(3), (2)+(3), and compare them with the classification performances when all three sets of features are used (1)+(2)+(3). The main idea is that if a set of features is particularly important for the classification, the performance of the classification algorithms should significantly drop when this set of features is excluded. All experiments were conducted using the 5-fold cross-validation with 10 repetitions in Weka Experimenter.

The results of these experiments for British English are presented in Table 3.26.

The results (classification performances) of each experiment were compared with the experiment in which all features were used (row 'all' in Table 3.26), using the twotailed paired t-test at a 0.05 level of significance provided by Weka Experimenter. Cases in which the differences between classifier performance of the particular experiment was significantly lower than in the experiment where all features were used are denoted by an '*'. There were no cases in which a classifier's accuracy in any experiment outperformed the accuracy of the same classifier in the experiment with all features.

Set	SMO(n)	SMO(s)	NB
all	90.19	92.52	85.48
(1)	68.81*	64.05*	67.86*
(2)	68.19*	70.14*	70.33*
(3)	87.24	92.67	88.81
(2)+(3)	91.14	93.33	86.48
(1)+(3)	87.38	89.52	88.43
(1)+(2)	74.48*	66.71*	70.33*

Table3.26: Diachronic classification in British English

3.7Conclusion :

In this chapter we studied 4 works related to our thesis. Their titles were as follows :

1 : On Predicting Religion Labels in Microblogging Networks

2 : Detection and classification of social media‐based extremist affiliations using sentiment analysis techniques

3 : Lexical Analysis of Religious Texts using Text Mining and Machine Learning Tools

4 : Style of Religious Texts in 20th Century

Where we mentioned the method and the most prominent algorithms used in their research. We presented the results obtained. Then we compared the results in each of the actions we mentioned.

Chapter4: The proposed approach

(Features, Dictionarys, Dataset)

4. The proposed approach (Features, Dictionarys,Dataset).

4.1 Introduction:

In this chapter, we start by identifying the features of each religion to obtain a more accurate data set and then we collected the data to which the form is applied. Finally, we built dictionaries.

Contribution

Our main contributions are:

A) Work on various religion (Islam, Christianity, Judaism, Atheism) by exploiting four compilations, support vector machines (SVM), decision tree (DT), random forest (RF) and naive Bayes (NB).

B) An explanation of the dataset containing 3373 comments.

C) Create a set of dictionaries for each religion (religious phrases, divine books, supplications, religious terms, religious hadiths, religious figures ..... etc.).

4.2 Proposed features:

The main thing that distinguishes religious texts is the use of some religious words and phrases and the cite of divine books.... Etc.

We have identified the features of each religion that we will explain bellow:

- **Feature of Islam text :**
  - Cite of Quranic verses and prophetic sayings.
  - Use of religious phrases : Alhamdulilla , Mashaa'Allah , Allahumma Salli Wa Sallam Wa bareek ala habibulLAH sallallahu Alayhy Wa Sallam, Allah Subhan Wa Taala , La Illaha Ilallah... Muhammad Rasulullah.... etc
  - Cite of the greatest personalities of Islamic history : Imam Al-Shafi'i , Imam Al-Bukhari ..... etc.
  - Cite of the names of the Prophets, Messengers and Companions : Muhammad, Adam , Abu Bakr Al-Siddiq, Othman bin Affan ..... etc
  - Citing the attributes of Allah
  - The use of the Hijri months: Shaaban. Ramadan. Shawwal ... etc
  - Using supplications in abundance

- **Feature of Christianity text :**
  - Citing the gospel.
  - Citing by the names of Christ: Holy Spirit , Jesus Christ, Heavenly Father, Lord Jesus, Lord Jesus Christ, the Church, Saviour, Jesus, Jesus, the Scriptures, Christianity, the Messiah, the Gosbel, the Spirit……. Etc.
  - Use of religious terms: Amen in Jesus Holy Name, Holy Spirit, Thank you Jesus Christ,Praise you in the highest, Lord Jesus…….. Etc.
  - The use of religious expressions: God the Father, Son Jesus, Holy Spirit, Son of God, the name of Jesus, in Jesus Holy Name…. Etc.
- **Feature of Judaism text :**
  - Citing the Bible.
  - The use of these phrases : glory and eternity for the Lord, Shabat Shalom, God bless Israel, God bless the Jews, I am a Jew, God stands with his children and We are God's chosen people.
  - Writing the words of God in this way "G-D."
  - Citing rabbis.
  - The use of the word Chadian, Jehovah, which means Lord.
- **Feature of Atheism text :**
  - To deny the existence of God.
  - Questioning the existence of God.
  - Using questioning tools.
  - Use of negation tools: I do not believe, do not think, I do not doubt.
  - Lying the Qur'an and not believing it.
  - Atheists use the phrase: "god of gaps", i am not religious, i disbelieve in all religions, when inventing a god, your alleged god, morality originates in atheism.

    . Citing the most famous name of atheists: Jason Aaron-Clark Adams-larry Adler-Amy Alkon-Robert Altman-Natalie Angier-Aziz Ansari-Isaac Asimov-Scott Atran-Bob Avakian.....

4.3 Dataset:

Annotation:

Collecting and classifying data is a rather difficult task because it takes a lot of time to classify it as one that pollates the other and needs a lot of knowledge to reach a final decision that defines the category of each comment (Muslim, Christian, Jewish, Atheist or neutral).

The latter took us between five and six weeks and two commentators (the third until the dispute between commentators is resolved by a majority vote) to collect 3373 comments from different religious groups on Facebook, among them Jewish Spirit, Christian Support Group, Islam for Muslim, Global Islamic Orator Group, the Athheists Group, the Athheist Alliance of America …….Etc. It was then classified into five categories: Islamic, Christianity, Jewish, Judaism and Neutral.

The following table represents data statistics that are determined by polar values.

Table 3.1 divides the Dataset according to the five polarity value:

Polarity	Islamic	Christianity	Judaism	Atheism	Neutral	Total
Number of comments	**701**	**643**	**716**	**706**	**607**	**3373**

Table 4.1: Number of comments by polarity.

In Table 4.2 we have described some examples with their polarities.

Comment	Polarity
If God exists why can't we see Him?	Atheist.
If God is fair then why are diseases spread and why is injustice permitted among people.	Atheist.
Subhun Allah Alhamduillah La ilaha illallah Allahu Akbar.	Muslim
Ameen InshaAllah i will meet my husband in jannah ul firdus Ameen.	Muslim
Amen thank you Jesus my Lord and Savior.	Christian
HALLELUJAH. GLORY TO GOD ALMIGHTY. THANK YOU JESUS CHRIST AMEN.	Christian
just Like The Bible Said Israel Will Turn The Desert Into Paradise Israel Jews,	Jewish
God belongs to the people of the Jews and not to others	Jewish
Please put your menu online!!	Neutral
Wow that really works for me thanks man.	Neutral

Table 4.2: An example of a classification of some comments.

We have worked on data in several cases in classes of the number, type of classes and size of data we will explain what we did in these points:

- Distinction between religious texts.
- Work on a balanced and unbalanced number of text.
- The distinction of religious texts is made in a variety of texts (religious and non-religious).
- Work on a different number of classes (five, four, two).

4.4Proposed Classifications

We have counted 18 cases, we will explain what we did at the following:

1. Work on only four religious classes (Islamic, Christian, Jewish, and Atheism).
2. Work on five classes (Islamic, Christian, Jewish, Atheism, neutral).
3. Working on two religious classes only:The first class includes only Islamic comments and the second class includes the rest of the religious

comments (Christian, Jewish and Atheism) and the number of comments in both classes is balanced.

4. Work on two religious classes only: the first class includes Christian comments only, and the second class includes the rest of the religious comments (Islamic, Judaism and Atheism) and the number of comments in both classes is balanced.
5. Work on two religious classes only: the first class includes Jewish comments only, and the second class includes the rest of the religious comments (Islamic, Christian, and Atheism) and the number of comments in both classes is balanced.
6. Working on two religious classes only: the first class includes Atheism comments only, and the second class includes the rest of the religious comments (Islamic, Christian, and Jewish) and the number of comments in both classes is balanced.
7. Working on two religious classes only: the first class includes Islamic comments only, and the second class includes the rest of the religious comments (Christian, Judaism and Atheism) and the number of comments in both the first and second classes is unbalanced.
8. Work on two religious classes only: the first class includes Christian comments only, and the second class includes the rest of the religious comments (Islamic, Judaism and Atheism) and the number of comments in both classes is unbalanced.
9. Work on two religious classes only: the first class includes Jewish comments only, and the second class includes the rest of the religious comments (Islamic, Christian, and Atheism) and the number of comments in both classes is unbalanced.
10. Work on two religious classes only: the first class includes Atheism comments only, and the second class includes the rest of the religious comments (Islamic, Christian, and Jewish) and the number of comments in both classes is unbalanced.
11. Work on two classes: the first class includes Islamic comments only, and the second class includes the rest of the religious and non-religious

comments (Christian, Judaism, Atheism and neutral) and the number of comments in both classes is balanced.

12. Work on two classes: the first class includes Christian comments only, and the second class includes the rest of the religious and non-religious comments (Islamic, Judaism, Atheism and neutral) and the number of comments in both classes is balanced.
13. Work on two classes: the first class includes Jewish comments only, and the second class includes the rest of the religious and non-religious comments (Islamic, Christian, Atheism and neutral) and the number of comments in both classes is balanced.
14. Work on two classes: the first class includes Atheism comments only, and the second class includes the rest of the religious and non-religious comments (Islamic, Christian, Jewish and neutral) and the number of comments in both classes is balanced.
15. Working on two classes: the first class includes Islamic comments only, and the second class includes the rest of the religious and non-religious comments (Christian, Judaism, Atheism and neutral) and the number of comments in each of the first and second classes is unbalanced.
16. Work on two classes: the first class includes Christian comments only, and the second class includes the rest of the religious and non-religious comments (Islamic, Judaism, Atheism and neutral) and the number of comments in both classes is unbalanced.
17. Work on two classes: the first class includes Jewish comments only, and the second class includes the rest of the religious and non-religious comments (Islamic, Christian, Atheism, and neutral) and the number of comments in both classes is unbalanced.
18. Work on two classes: the first class includes Atheism comments only, and the second class includes the rest of the religious and non-religious comments (Islamic, Christian, Jewish and neutral) and the number of comments in both classes is unbalanced.

4.5 Create the dictionary

It is a set of words and sentences that a workbook can use to evaluate the polarity of the text to obtain a more accurate classification and reduce margin of error.

We construct our own dictionary based on what we found in the comments and on the features and , each dictionary is a set of dictionaries. It took nearly three weeks of work to capture a large group of words and sentences, which are shown in Table 4.3:

Polarity	Islamic	Christianity	Judaism	Atheism	Total
Number of words and sentences	7389	132	30179	1156	38856

Table4.3 : Our dictionary's statistics.

❖ Islamic Dictionary:

Name of dictionary	A_ Allah	D_ Muslim	H_ Hadiths	H_ months	P_ Muslim	Py_ Muslim	Quran	S_ Quran	Total
Number of words and sentences	99	570	116	13	87	43	6347	114	7389

Table4.4: Statistics of islamic dictionary.

Table 4.6 and Table 4.5 represent a sample of Islamic dictionary

A_Allah	D_mulim	H_hadiths
Allah	Surah	*Abu Hurairah (RA) said: The Messenger of Allah (peace be upon him) said, "You must not supplicate: 'O Allah! Forgive me if You wish; O Allah bestow mercy on me if You wish.' But beg from Allah with certitude for no one has the power to compel Allah." (Riyad as-Salihin, Book 18, Hadith 1743)*
Ar-Rahmaan	Alhamdulilla	
Ar-Raheem	Inshaallah	
Al-Malik	Mashaa'Allah	
Al-Quddoos	In sha Allah	
As-Salaam	Masha Allah	
Al-Mu'min	May God guide us	
Al-Muhaimin	May ALLAH guide us	
Al-Azeez	Allahumma ameen	
Al-Jabbaar	subhanAllah	
Al-Mutakabbir	Ya Rabb	
Al-Khaaliq	AMEEN	
Al-Bari'	Mashaallah	
Al-Musawwir	Barakallah	
Al-Ghaffaar	Subhanallah	
Al-Qahhaar	Assalamualikum	
Al-Wahhaab	ALLAH	
Al-Razzaaq	Amin	
Al-Fattaah	Allah Subhan Wa Taala	
Al-Aleem	S.A.W	
Al-Qaabid	adhkaar	
Al-Baasit	Sunnah	
Al-Khaafid	prophet Muhammad	
Ar-Raafi	Peace be upon him	
Al-Muiz	Assalamu aleikum ua rahmutallahi!	
Al-Muthil	Subuhana Ta'ala	
As-Samee	Prayer	
Al-Baseer	Fasting	
Al-Hakam	Quran	
		Abu Harairah (RA) narrated that the Messenger of Allah (peace be upon him) said: "Three supplications are accepted , there is no doubt in them (about them being accepted): The supplication of the oppressed, the supplication of the traveler, and the supplication of his father against his son." (Jami`at-Tirmidhi, Book 27, Hadith 11)
		It was narrated from Jabir bin 'Abdullah (RA) that the Messenger of Allah (peace be upon him) said: "O people, fear Allah and be moderate in seeking a living, for no soul will die until it has received all its provision, even if it is slow in coming. So fear Allah and be moderate in seeking provision; take that which is permissible and leave that which is forbidden." (Sunan Ibn Majah, Book 12, Hadith 2227)

Table 4.5: Sample of Islamic dictionary (part1) .

H_months	P_Muslim	Py_Muslim	Quran	S_Quran
muharram	Lady Fathimah	Morning prayer	From the mischief of Darkness as it overspreads;	Al Fatihah
safar	Imam Ali	Aduher	From the mischief of those who practise secret arts;	Al-Baqarah
rabi al-awwal	Adam	Asr	And from the mischief of the envious one as he practises envy.	Al'Imran
rabi al-thani	Idris	Maghreb	In the name of Allah, Most Gracious, Most Merciful.	AnNisa
jumada al-awwal	Nuh	eshaa	Say: I seek refuge with the Lord and Cherisher of Mankind,	AlMa'idah
jumada al-thani	Hud	Duha prayer	The King (or Ruler) of Mankind,	AlAn'am
rajab	Saleh	Tahajud	The god (or judge) of Mankind,-	AlA'raf
sha'ban	Ibrahim	Prayers for rain	From the mischief of the Whisperer (of Evil), who withdraws (after his whisper),-	AlAnfal
Ramadan	Lut	Prayer for Guidance	(The same) who whispers into the hearts of Mankind,-	AtTaubah
ramadan	Ismail	Voluntary Prayer	Among Jinns and among men.	Yunus
shawwal	Ishaq	Optional Night Prayer		Hud
dhul qa'dah	Yaqub	Prayer of Penitence		Yusuf
dhul hijjah	Yusuf	Congregational Prayer		ArRa'd
	Ayyub	Prayer of Need		Ibrahim
	Shu?ayb	Prayer of Fear		AlHijr
	Musa	Prayer during Journey		AnNahl
	Harun	Supererogatory Prayer		AlIsra
	Dhul-Kifl	Morning Prayer		AlKahf
	Dawud	Forenoon Prayer		Maryam
	Sulayman	Noon Prayer		TaHa
	Ilyas	Prayers of the Two Feasts		AlAnbiya
	Al-Yasa			AlHajj
				AlMu'minun
				AnNur
				AlFurqan

Table 4.6: Sample of Islamic dictionary (part2) .

❖ Christianity Dictionary

Name of dictionary	D_christ	Total
Number of words and sentences	132	132

Table4.7 : Statistics of christianity dictionary.

Table 4.8 represent a sample of christianity dictionary.

D_christ
Holy Spirit
Jesus Christ
Heavenly Father
Lord Jesus
Lord Jesus Christ
the church
Saviour
Jesus
son Jesus
Holy Spirit
son of God
yeshua
the Scriptures
the Psalms
Christianity
The messiah
AMEN JESUS CHRIST
" Render therefore to all their dues: tribute whom tribute is due custom to whom custom fear to whom fear honour to whom honour." (Romans 13:7) KJV
Jesus said, “He that believes and is plunged into water shall be saved…” (Mark 16:15,16)
Our ALMIGHTY FATHER OUR LORD JESUS CHRIST His the HEALER

Table 4.8: A sample of christianity dictionary.

- **Judaism dictionary**

Name of dictionary	Bible	F_Jews	N_famous_jews	S_jews	Total
Number of words and sentences	29158	685	194	142	30179

Table4.9 : Statistics of judaism dictionary.

Table 4.10 represent a sample of judaism dictionary

Bible	*F_Jews*	*N_famous_jews*	*S_jews*
In the beginning God created the heavens and the earth.	I'm proud to stand up for the Israelis."	The most prominent flags of the Jews	The most prominent religious phrases
Now the earth was formless and empty. Darkness was on the surface of the deep. God's Spirit was hovering over the surface of the waters.	Amen Israel is Gods Holy Land	Albert Einstein	And write them on your house door lists and on your doors
God said, "Let there be light," and there was light.	Call to His Name Yahawah .	Abraham Limbel	"The children of the settlers who have come down with you shall enslave them forever ... and take slaves and slaves from them ... your brothers from the children of Israel, and no man overthrows his brother violently" (Leviticus).
God saw the light, and saw that it was good. God divided the light from the darkness.	God of Israe	Arthur Pearson	
God called the light "day," and the darkness he called "night." There was evening and there was morning, one day.	God of jewish	Arthur Korn	
God said, "Let there be an expanse in the middle of the waters, and let it divide the waters from the waters."	May God send His choicest blessings in Israel	Aaron Aaronson	
God made the expanse, and divided the waters which were under the expanse from the waters which were above the expanse and it was so.	the enemy of Israel can defeated because god never forget his choosen people,	Asher Pires	
	Only god the father can defend and protect Israel,	Abram Yovi	Now I have seen this. So, I can tell you that this is God's Son.'
	Israel is a Hero from GOD	Abraham Smalovich Psekovovich	John was standing there again the next day, with two of his disciples
	You see God of Israel so powerful that help them won their enemies.	Adolf Lowe	
	God is with His chosen people always and He meant it when He said " I will bless those who bless you and curse those who curse you to Israel "	Albert Neuberger	Glory and eternity to the Lord
		Alfred Adler	We are God's chosen people
		Alexander Ostrovsky	Israel is the state of truth
		Alexander Luria	Long live the Jewish people
		Alexandra Adler	
		Amnon Yariv	
		Amir Bennouli	
		Anatole Abragam	
		Andre Loew	
		Otto Farburg	
		Erna Foreman	
		Ernest Gellner	
		Ernest Kris	
		Erwin Chargaf	
		Eric From	
		Eric Candle	
		Ishaq bin Syed	
		Ephraim Katzir	

Table 4.10: A sample of judaism dictionary.

❖ Atheism dictionary

Name of dictionary	A_Atheist	NF_atheist	Total
Number of words and sentences	773	383	1156

Table 4.11: Statistics of atheism dictionary.

Table 4.12 represent a sample of atheism dictionary.

A_Atheist	*NF_atheist*
Isnt not believing in god the whole definition of atheist lol	Atheists
I don't 'believe' there is no such thing as god.I KNOW it.	Zaki Arsuzi
If God exists why can't we see Him?	John Anderson
Atheism is a more philosophical decision than a science-based decision a brave decision that may require atoms of faith that there is no God.	Louise Anthony
if god really exist Why do not accept supplications at the same time?	Alfred Air
On what verses do you speak of the seven lands that your Qur'an claims are that God created them or that the sun runs to a stable for it or does it sink into a mud pool Sure you are not educated.	Julian Baggeni
Your God is the one leading astray.	Mikhail Bakunin
Do you think that this universe will be left with its galaxies stars and everything that created it to pray to us?.	Roland Bart
Is it reasonable for the Creator of this universe who created us to worship to pray to the servant of his creation and a prophet who sent Him to us?.	George Batai
For God and life is matter.	Bruno Power
I feel psychological relief when I remember that I disbelieved in God while I was young ..aaah how long would I suffer if I disbelieved after a long life?.	Jan Baudrillard
	Simone de Beauvoir
	Jeremy Bentham
	Simon Blackburn
	Peter Boghossian
	Martin Bodry
	Celestin Bogley
	Gustavo Bueno
	David Chalmers
	Diprasad Chatubadaya
	Patricia Churchland
	Paul Churchland
	Marques de Condorcet
	Auguste Comte

Table 4.12: A sample of atheism dictionary

4.6 Conclusion:

We did a lot of work in this chapter, identifying general and specific hints for each religious class and explaining a data set of 3373 texts, which were classified as five categories (Islamic, Christian, Jewish, atheist, and neutral). Finally, we created a dictionary of 38,856 words and phrases.

Chapter 5: Implementation and Results.

5. Implementation and Results

5.1 Introduction

In this chapter, we will implement religious detection algorithms. We will mainly implement four algorithms. These algorithms are: Support Vector Machines (SVM), Tree Decision (DT), Random Forest (RF) and Naïve Bayes (NB). We'll use a corpus ready to use. And we will analyze and discuss the various classification results obtained after applying the four algorithms on the set of classifications previously proposed.

5.2Implementation steps

- The declaration of the libraries, which will be worked on during the programming process.

```
# In[1]:
#library declaration
import csv
from   operator import add
import pandas as pd
from   sklearn.svm import SVC
from   sklearn.externals import joblib
from   sklearn.model_selection import train_test_split
from   sklearn.tree import DecisionTreeClassifier
from   sklearn.ensemble import RandomForestClassifier
from   sklearn.naive_bayes import MultinomialNB
from   sklearn.metrics import accuracy_score
from   sklearn.metrics import precision_recall_fscore_support
```

Figure 5.1: Declaration of the libraries

- In this section, the dataset (all comments) and all features are read and listed in tables to be handled later

```
# In[3]:
# Read comment data set and all the existing features

# All Data set
Corpus = pd.read_csv(r"E:/Master20/Dataset/DS_CSV/Dataset.csv" ,encoding = 'unicode_escape' , engine='python' )

# Features of the religion of Islam
M_Quran      = pd.read_csv(r"E:/Master20/Dataset/Muslim/Quran.csv"     , header=0, sep='delimiter')
M_hadith     = pd.read_csv(r"E:/Master20/Dataset/Muslim/H_hadiths.csv" , header=0, sep='delimiter')
M_muslim     = pd.read_csv(r"E:/Master20/Dataset/Muslim/D_muslim.csv"  , header=0, sep='delimiter')
M_Allah      = pd.read_csv(r"E:/Master20/Dataset/Muslim/A_Allah.csv"   , header=0, sep='delimiter')
M_P_muslim   = pd.read_csv(r"E:/Master20/Dataset/Muslim/P_Muslim.csv"  , header=0, sep='delimiter')
M_Py_muslim  = pd.read_csv(r"E:/Master20/Dataset/Muslim/Py_Muslim.csv" , header=0, sep='delimiter')
M_S_Quran    = pd.read_csv(r"E:/Master20/Dataset/Muslim/S_Quran.csv"   , header=0, sep='delimiter')
M_month      = pd.read_csv(r"E:/Master20/Dataset/Muslim/H_months.csv"  , header=0, sep='delimiter')

# Features of the Christian religions
C_christ     = pd.read_csv(r"E:/Master20/Dataset/Christ/D_christ.csv"  , header=0, sep='delimiter')


# Features of the Jewish Religion
J_Bible      = pd.read_csv(r"E:/Master20/Dataset/Jews/Bible.csv"       , header=0, sep='delimiter')
J_Feature    = pd.read_csv(r"E:/Master20/Dataset/Jews/F_Jews.csv"      , header=0, sep='delimiter')
J_N_famous   =pd.read_csv(r"E:/Master20/Dataset/Jews/N_famous_jews.csv", header=0, sep='delimiter')
J_Sentence   = pd.read_csv(r"E:/Master20/Dataset/Jews/S_jews.csv"     , header=0, sep='delimiter')

# Features of an atheist
A_Atheists   = pd.read_csv(r"E:/Master20/Dataset/Atheist/A_Atheist.csv" , header=0, sep='delimiter')
A_NaaaF      = pd.read_csv(r"E:/Master20/Dataset/Atheist/NF_atheist.csv", header=0, sep='delimiter')
```

Figure 5.2: Read and listed dataset and features.

- First: The program counts the words in each comment of the Dataset (We must recall that the separation between the word and the word is the space and the Split function () are the ones that perform this operation).

```
# In[4]
# Calculate the number of words in each comment from data set
word_count = []
for i in Corpus.COMMENT:
    word_count.append(len(i.split()))

Corpus["word_count"] = word_count
```

Figure 5.3: Counts the words in each comment of the Dataset.

the result is a **Word_count** table that shows the word count for each comment and then lists the result in a column called **Word_count** in the Corpus table .

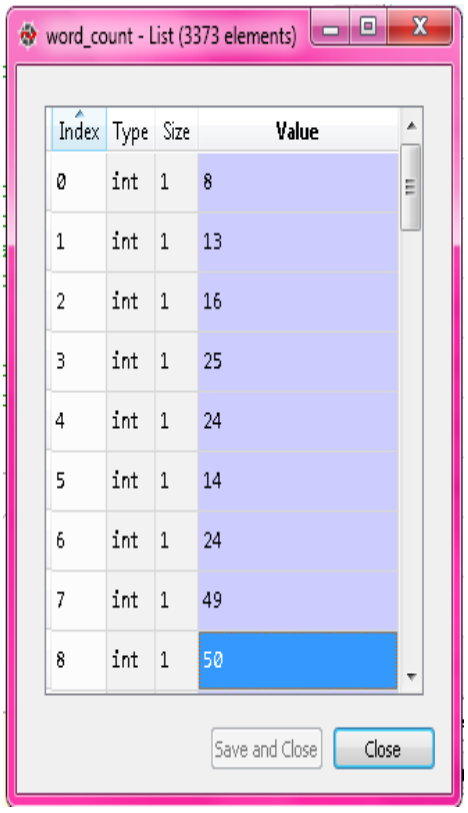


Figure 5.4: Word count .

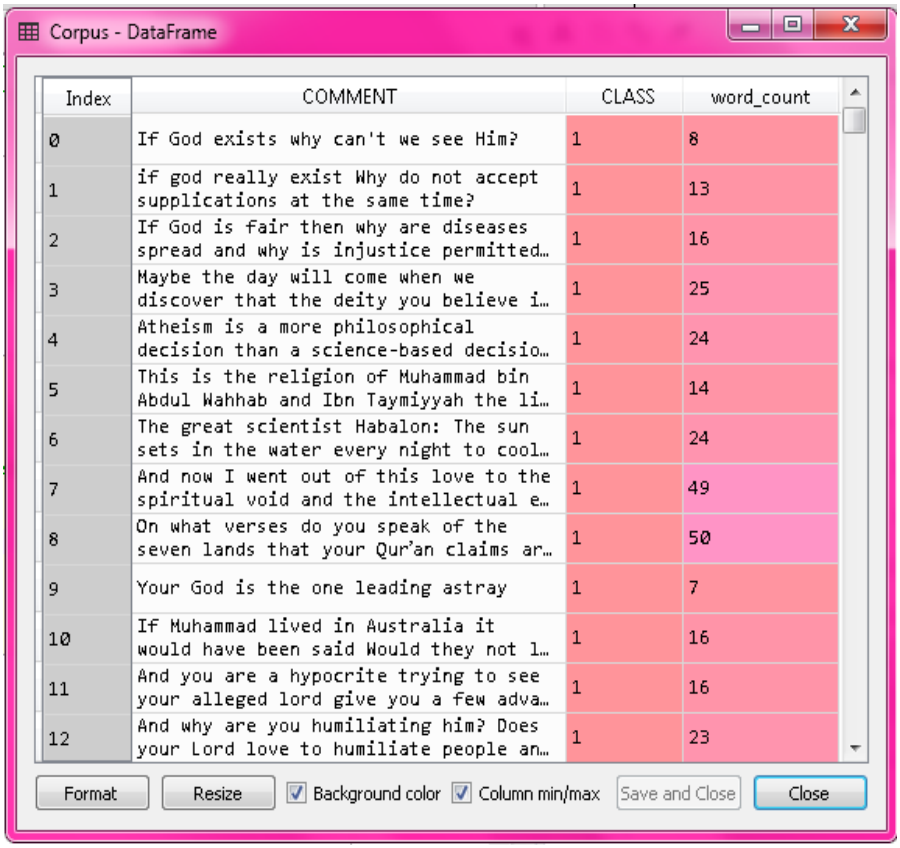


Figure 5.5: Data after adding a column of word count.

- Second: The program creates a table with vectors for comments so that it separates each comment for a unit and here it represents a vector and separates each word of the comment separately to become a component of the vectors.

```
# In[ 5]:

Corpus_words      =[]

csv_reader = csv.reader(open(r'E:/Master20/Dataset/DS_CSV/Comment.csv'  ), delimiter=',')
arr = list(csv_reader)
for line in arr:
        Corpus_words.append(line[0].split(" "))
```

Figure 5.6: Represent comments and words in vectors.

The results :

Figure 5.7: The Results of representations of comments and words in vectors

We perform this process with all our feature tables Some exemples:

Figure 5.8: Example 1 of representing comments and words in vectors.

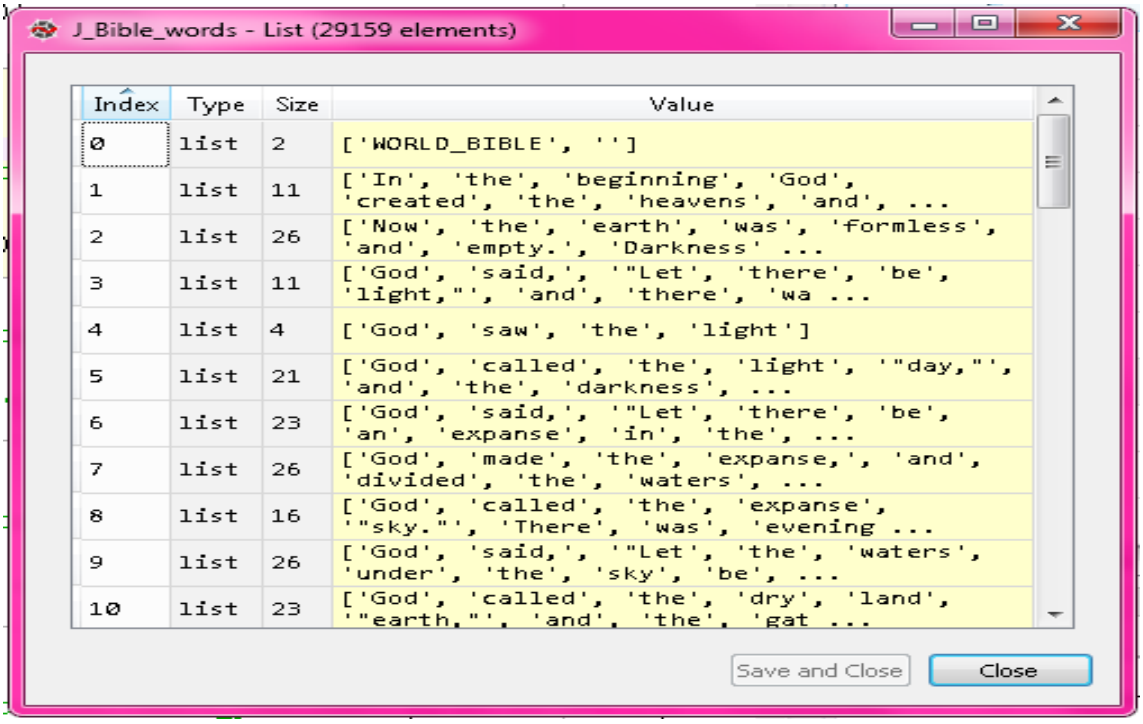


Figure 5.9: Example2 of representing comments and words in vectors.

- Third: The program scans all the words of the datasat on the tables of the features table in a table for each vector separately (comment) If there is a match that gives the value 1 or the value 0, then insert these values into a column in the dataset table (Corpus), in our example this column name in the table one_P_muslim.

```
247
248 F_P_muslim   = []
249 for t in Corpus.COMMENT:
250     if any(w in t.split() for w in M_P_muslim.text):
251         F_P_muslim.append(1)
252     else :
253         F_P_muslim.append(0)
254
255 Corpus["one_P_muslim"] = F_P_muslim
```

Figure 5.10: Scans all the words of the datasat on the tables of the features

The results :

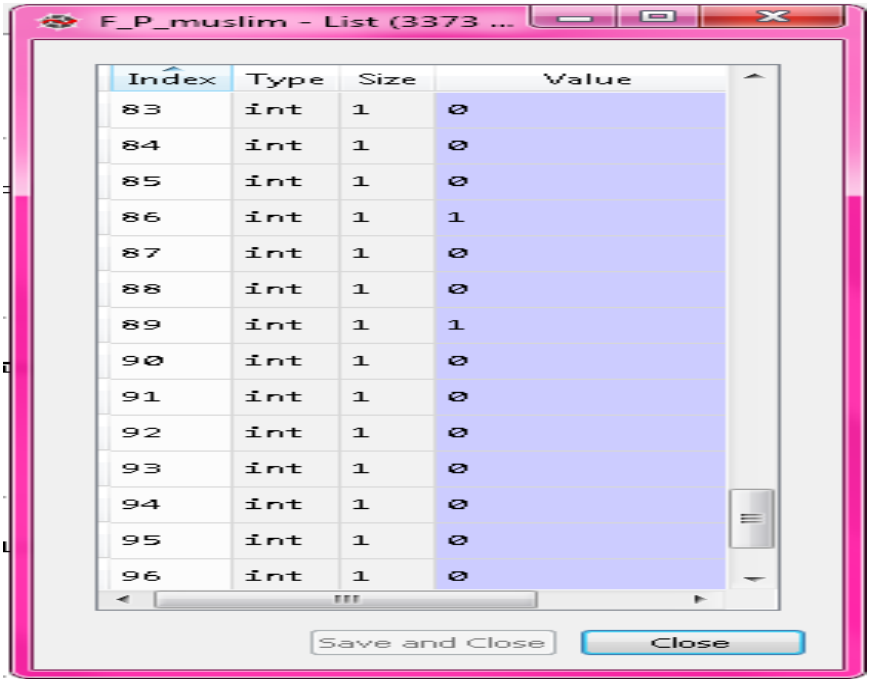


Figure 5.11: Result of scanning .

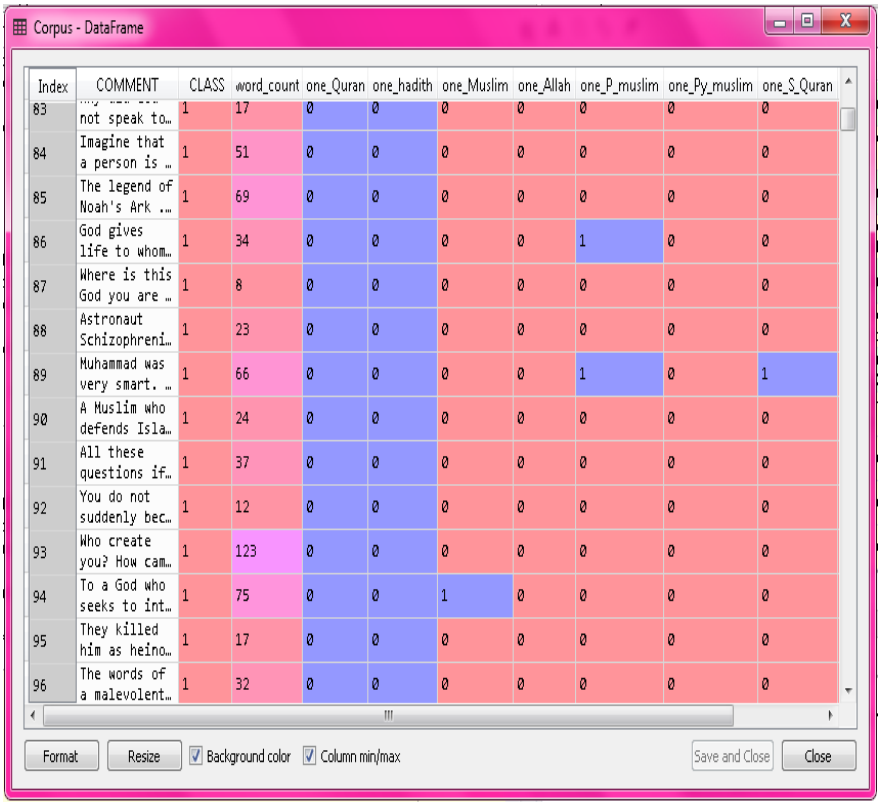


Figure 5.12: Insertion the Match values into dataset (Corpus)

A number 1 in blue is the result of a match in that line with the attributes in the column header.

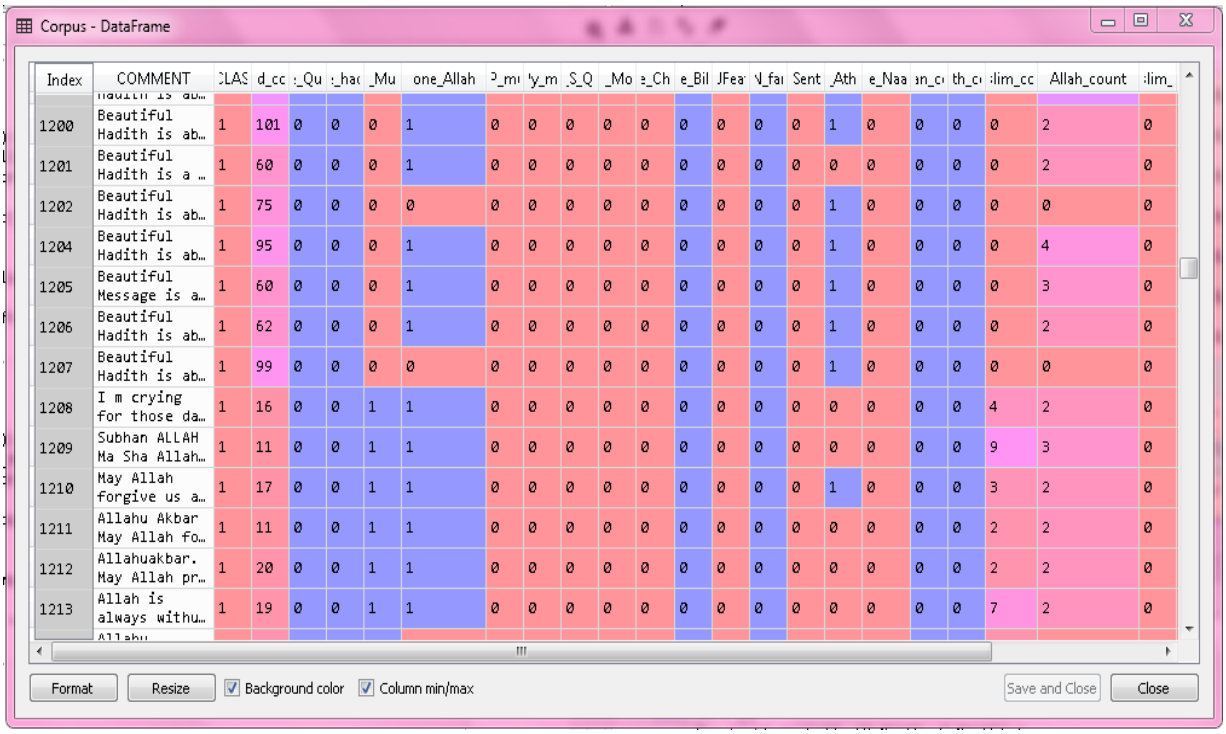
- Fourth: After we find matches with features we collect them; (collect them for each individual feature), this process with all the table of features.We insert the result in columns in the dataset table(corpus), for example here we inserted under the name (Quran_count).

Note: Features are often complete sentences.

```
365 # In[18]:
366
367 CW_Quran       = []
368 count = 0
369 for t in Corpus.COMMENT:
370     for word in t.split():
371         for word1 in M_Quran.text:
372             if word == word1:
373                 count +=1
374     CW_Quran.append(count)
375     count = 0
376 fn = []
377 fn= list( map(add, CW_Quran, F_Quran) )
378
379 Corpus["Quran_count"] = fn
```

Figure 5.13: Collect matches with features

The results:



Index	COMMENT	CLAS	d_cc	:_Qu	:_hac	_Mu	one_Allah	?_mu	'y_m	_S_Q	_Mo	e_Ch	e_Bil	JFea	N_fai	Sent	_Ath	e_Naa	an_c	th_c	:lim_cc	Allah_count	:lim_
1200	Beautiful Hadith is ab…	1	101	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	2	0
1201	Beautiful Hadith is a …	1	60	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
1202	Beautiful Hadith is ab…	1	75	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
1204	Beautiful Hadith is ab…	1	95	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	4	0
1205	Beautiful Message is a…	1	60	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	3	0
1206	Beautiful Hadith is ab…	1	62	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	2	0
1207	Beautiful Hadith is ab…	1	99	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
1208	I m crying for those da…	1	16	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	4	2	0
1209	Subhan ALLAH Ma Sha Allah…	1	11	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	9	3	0
1210	May Allah forgive us a…	1	17	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	3	2	0
1211	Allahu Akbar May Allah fo…	1	11	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0
1212	Allahuakbar. May Allah pr…	1	20	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0
1213	Allah is always withu…	1	19	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	7	2	0

Figure 5.14: Results of collating matches.

- **Machine Training and Results draw phase:**

First: We take the values that the system will train on and take the test

```
# In[23]:


Z = Corpus.get(Corpus.columns[1:])
Y = Corpus.get(Z.columns[0])
X = Corpus.get(Z.columns[1:])
x_train, x_test, y_train, y_test = train_test_split(X, Y, test_size=0.2, random_state=4)
```

Figure 5.15: Preparation the corpus for training and test

The **get** function gives only numerical values from these corpus ,These are X values, and Y values are the class values.

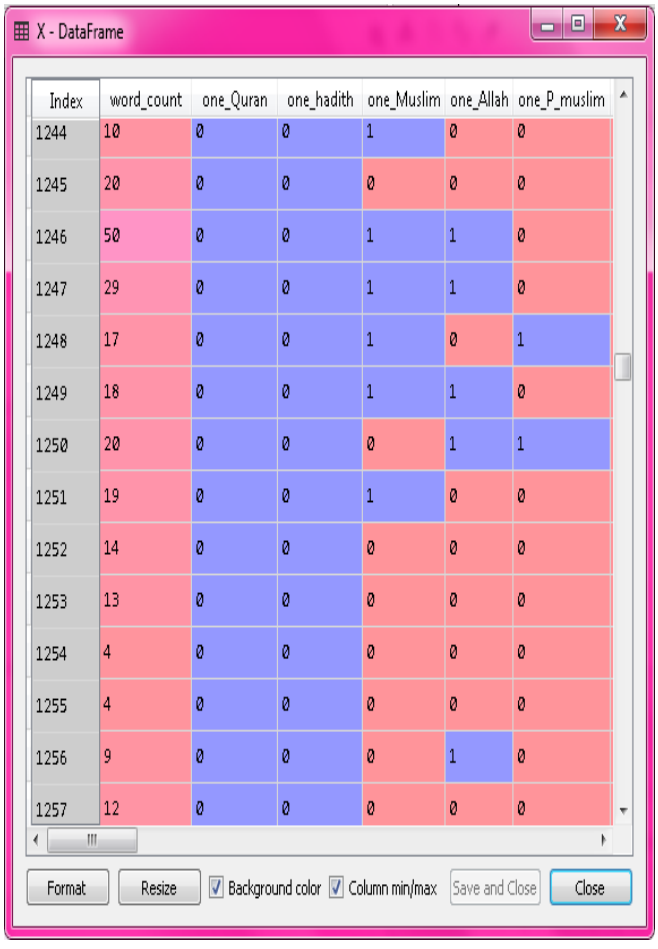


Figure 5.16: Numerical values of corpus.

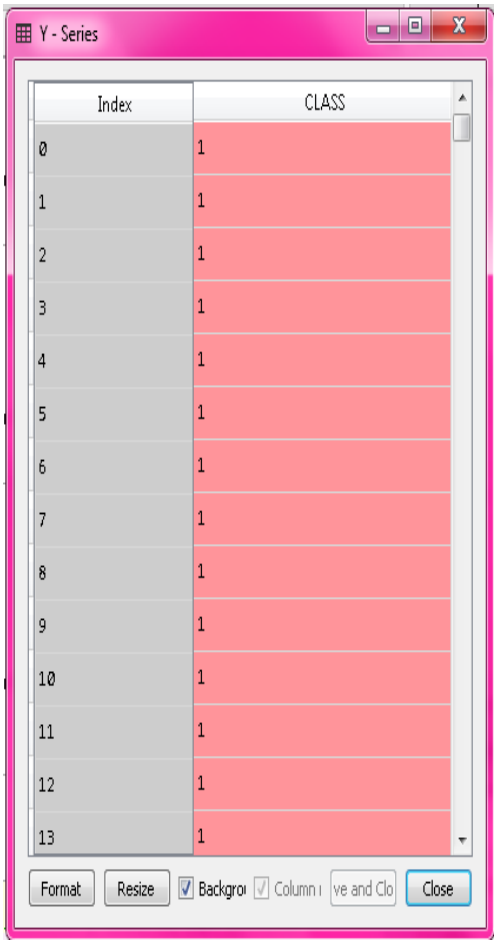


Figure 5.17: The class values

Then we add them to the value of Z.

Figure 5.18: Numerical values and classes values of corpus.

- And then we pass the data on the four algorithms as follows to create the functions required from each algorithm.

```
# In[24]:

svm = SVC()
svm.fit(x_train, y_train)
acc1_svm = svm.score(x_test, y_test)
pr1_svm = svm.predict(x_test)
ac = accuracy_score(y_test, pr1_svm)
ww1 = precision_recall_fscore_support(y_test, pr1_svm, average='macro')


# In[25]:

dt = DecisionTreeClassifier()
dt.fit(x_train, y_train)
acc2_dt = dt.score(x_test, y_test)
pr2_dt = dt.predict(x_test)
ww2 = precision_recall_fscore_support(y_test, pr2_dt, average='macro')
ac2 = accuracy_score(y_test, pr2_dt)

# In[26]:

rf = RandomForestClassifier(n_estimators=40)
rf.fit(x_train, y_train)
acc3_rf = rf.score(x_test, y_test)
pr3_rf = rf.predict(x_test)
ww3 = precision_recall_fscore_support(y_test, pr3_rf, average='macro')
ac3 = accuracy_score(y_test, pr3_rf)

# In[27]:

NB = MultinomialNB()
NB.fit(x_train, y_train)
acc4_NB = NB.score(x_test, y_test)
pr4_NB = NB.predict(x_test)
ac4 = accuracy_score(y_test, pr4_NB)
ww4 = precision_recall_fscore_support(y_test, pr4_NB, average='macro')
```

Figure 5.19: Practice x and y and calculate accuracy.

- Finally, let's get the accuracy of each algorithm.

Name	Type	Size	Value
ac	float64	1	0.9288888888888889
ac2	float64	1	0.9244444444444444
ac3	float64	1	0.9333333333333333
ac4	float64	1	0.9111111111111111
acc1_svm	float64	1	0.9288888888888889
acc2_dt	float64	1	0.9244444444444444
acc3_rf	float64	1	0.9333333333333333
acc4_NB	float64	1	0.9111111111111111

Figure 5.20: The accuracy of each algorithm

Note that RF is the best algorithm so we keep the model.

```
660 # In[28]:
661     # Output a pickle file for the model
662 joblib.dump(svm, 'Model_of_Christian.pkl')
663
```

Figure 5.21: Instructions for saving the modal

```
In [28]: joblib.dump(svm, 'Model_of_Christian.pkl')
Out[28]: ['Model_of_Christian.pkl']
```

Figure 5.22: Result of saving the modal

5.3 Experiments and results:

As long as we use the supervised learning method, we split the corpus into two parts, 80% for training and 20% for testing.

5.3.1 Results analysis:

We have performed several tests. table 5.1 shows the results of the classification and partial discussion of each test.

N° Test	Test or case	Classifier	Precision (%)	Discussion
1	two classes: the first class includes Islamic comments only, and the second class includes the rest of the religious and non-religious comments (Christian, Judaism, Atheism and neutral) and the number of comments in both classes is **(C1:700 ,C2:700).**	**SVM**	**90.03**	The best result in this test is **90.03%** given by the **SVM** and the worst result of this test is **88.25%** given by the **NB.**
		DT	**89.32**	
		RF	**89.67**	
		NB	**88.25**	
2	two religious classes only: the first class includes Christian comments only, and the second class includes the rest of the religious comments (Islamic, Judaism and Atheism) and the number of comments in both classes is **(C1:643 ,C2:643).**	**SVM**	**86.04**	The best result in this test is **86.04%** given by the **SVM** and **NB** the worst result of this test is **84.49%** given by the **RF**
		DT	**85.65**	
		RF	**84.49**	
		NB	**86 .01**	
3	two classes: the first class includes Christian comments only, and the second class includes the rest of the religious and non-religious comments (Islamic, Judaism, Atheism and neutral) and the number of comments in both classes is **(C1:643 ,C2:643).**	**SVM**	**87.54**	The best result in this test is **88.32%** given by the **NB** and the worst result of this test is **82.87%** given by the **DT.**
		DT	**82.87**	
		RF	**84.04**	
		NB	**88.32**	
4	Two religious classes only:The first class includes only Islamic comments and the second class includes the rest of the religious comments (Christian, Jewish and Atheism) and the number of comments in both classes is **(C1:700 ,C2:700).**	**SVM**	**86 .59**	The best result in this test is **86.59%** given by the **SVM** and the worst result of this test is **82.24%** given by the **RF**.
		DT	**84.05**	
		RF	**82. 24**	
		NB	**85.86**	

5	two religious classes only: the first class includes Atheism comments only, and the second class includes the rest of the religious comments (Islamic, Christian, and Jewish) and the number of comments in both classes is **(C1:706 ,C2:706).**	**SVM**	**75.97**	The best result in this test is **78.09%** given by the **NB** and the worst result of this test is **74.20%** given by the **RF.**
		DT	**75.26**	
		RF	**74.20**	
		NB	**78.09**	
6	. two classes: the first class includes Atheism comments only, and the second class includes the rest of the religious and non-religious comments (Islamic, Christian, Jewish and neutral) and the number of comments in both classes is **(C1:706 ,C2:706).**	**SVM**	**68.79**	The best result in this test is **68.79 %** given by the **SVM** and the worst result of this test is **66.31%** given by the **DT.**
		DT	**66.31**	
		RF	**67.02**	
		NB	**68.43**	
7	two classes: the first class includes Jewish comments only, and the second class includes the rest of the religious and non-religious comments (Islamic, Christian, Atheism and neutral) and the number of comments in both classes is **(C1:716,C2:716).**	**SVM**	**69.47**	The best result in this test is **69.47%** given by the **SVM** and the worst result of this test is **60%** given by the **NB.**
		DT	**68.42**	
		RF	**67.01**	
		NB	**60**	
8	two religious classes only: the first class includes Jewish comments only, and the second class includes the rest of the religious comments (Islamic, Christian, and Atheism) and the number of comments in both classes is **(C1:716 ,C2:716).**	**SVM**	**58.78**	The best result in this test is **58.78%** given by the **SVM** and the worst result of this test is **54.83%** given by the **DT.**
		DT	**54.83**	
		RF	**57.70**	
		NB	**58.06**	
9	two classes: the first class includes Islamic comments only, and the second class includes the rest of the religious and non-religious comments (Christian, Judaism, Atheism and neutral) and the number of comments in each of the first and second classes is**(C1:700 ,C2:2673)**.	**SVM**	**93.33**	The best result in this test is **93.77%** given by the **RF** and the worst result of this test is **90.96%** given by the **NB.**
		DT	**92.88**	
		RF	**93.77**	
		NB	**90.96**	
10	two classes: the first class includes Christian comments only, and the second class includes the rest of the religious and non-religious comments (Islamic, Judaism, Atheism and neutral) and the number of comments in both classes is **(C1:643 ,C2:2730).**	**SVM**	**92.88**	The best result in this test is **93.33%** given by the **RF** and the worst result of this test is **91.11%** given by the **NB.**
		DT	**92.44**	
		RF	**93.33**	
		NB	**91.11**	

11	two religious classes only: the first class includes Islamic comments only, and the second class includes the rest of the religious comments (Christian, Judaism and Atheism) and the number of comments in both the first and second classes is **(C1:700 ,C2:2065).**	**SVM**	**91.33**	The best result in this test is **91.33%** given by the **SVM** and the worst result of this test is **89.71%** given by the **DT.**
		DT	**89.71**	
		RF	**90.79**	
		NB	**90.07**	
12	two religious classes only: the first class includes Christian comments only, and the second class includes the rest of the religious comments (Islamic, Judaism and Atheism) and the number of comments in both classes is **(C1:643 ,C2:2122).**	**SVM**	**81.94**	The best result in this test is **81.94%** given by the **SVM** and the worst result of this test is **74.18%** given by the **DT.**
		DT	**74.18**	
		RF	**76.17**	
		NB	**78.70**	
13	two classes: the first class includes Jewish comments only, and the second class includes the rest of the religious and non-religious comments (Islamic, Christian, Atheism, and neutral) and the number of comments in both classes is **(C1:716 ,C2:2657).**	**SVM**	**80.14**	The best result in this test is **80.14%** given by the **SVM** and the worst result of this test is **66.96%** given by the **NB.**
		DT	**77.62**	
		RF	**78.66**	
		NB	**66.96**	
14	two classes: the first class includes Atheism comments only, and the second class includes the rest of the religious and non-religious comments (Islamic, Christian, Jewish and neutral) and the number of comments in both classes is **(C1:706 ,C2:2667).**	**SVM**	**78.37**	The best result in this test is **78.37%** given by the **SVM** and the worst result of this test is **70.51%** given by the **NB.**
		DT	**75.70**	
		RF	**76.44**	
		NB	**70.51**	
15	two religious classes only: the first class includes Jewish comments only, and the second class includes the rest of the religious comments (Islamic, Christian, and Atheism) and the number of comments in both classes is **(C1:716 ,C2:2049).**	**SVM**	**67.87**	The best result in this test is **67.87%** given by the **SVM** and the worst result of this test is **58.12%** given by the **NB.**
		DT	**60.64**	
		RF	**60.46**	
		NB	**58 .12**	

16	two religious classes only: the first class includes Atheism comments only, and the second class includes the rest of the religious comments (Islamic, Christian, and Jewish) and the number of comments in both classes is **(C1:706 ,C2:2059).**	**SVM**	**64.98**	The best result in this test is **64.98%** given by the **SVM** and the worst result of this test is **55.95%** given by the **NB.**
		DT	**56.85**	
		RF	**58.48**	
		NB	**55.95**	
17	Five classes (Islamic, Christian, Jewish, Atheism, neutral) and the number of comments in each class is **(C1:706 ,C2:700, C3:643 ,C4:716, C5:608).**	**SVM**	**55.40**	The best result in this test is **56.44%** given by the **RF** and the worst result of this test is **46.81%** given by the **NB.**
		DT	**54.07**	
		RF	**56.44**	
		NB	**46.81**	
18	Four classes :Islamic, Christian, Jewish, and Atheism, and the number of comments in each class is **(C1:706 ,C2:700, C3:643 ,C4:716).**	**SVM**	**53.79**	The best result in this test is **53.79%** given by the **SVM** and the worst result of this test is **46.57%** given by the **RF**
		DT	**49.81**	
		RF	**46.57**	
		NB	**53.61**	

Table 5.1: Classification results.

5.3.2 Results discussion:

- The best result in all tests is 93.77%, it gave by the RF with test 9 (two classes:The first class includes Islamic comments only, and the second class includes the rest of the religious and non-religious comments (Christian, Judaism, Atheism and neutral) and the number of comments in each of the first and second classes is unbalanced) , this is the same for SVM and DT, their best results were with the ninth test (93.33%, 92.88%respectively), but NB reached its maximum measurement (91.11%) in test (10).
- The worst result in all tests is 46.57%, it gave by the RF with test 18 (Four classes: Islamic, Christian, Jewish, and Atheism), this is the same for SVM

,DT , their worst results were with the test (53.79%, 49.81%respectively) , but NB reached its minimum measurement (46.81%) in the test (17).

- In tests (1,2,4,6,7,8,11,12,13,14,15,16,18) the best result for each of these tests presented by SVM. As for the tests (9, 10, 17) the best result for each of these tests presented by RF. And the rest of the tests (3 , 5) the best result for each of these tests presented by NB.
  - ❖ These results show that SVM is generally regarded as a better classifier.
- In tests (17) and (18) we found that the addition of class of neutral comments improved the results of classifiers except that NB which decreased the results. On the other hand, they had the worst results compared to the results of the rest of the tests (two classes).
  - ❖ Therefore, we can say that increasing the size of data has a greater impact on improving the result than increasing the number of classes (i.e. the less the number of classes and the greater the size of data, the better the result).
- In tests (4) and (11), (8) and (15), we found that Increasing the number of comments to the functionality of (test 4 and test 8) improved the results of all classifiers. In contrast to tests (2) and (12), (5) and (16) we found that the increase in the number of comments to the functionality of (test 2 and test 5) reduced the results of all the classifants.
- In tests (1) and (9), (3) and (10) , (7) and (13) , (6) and (14) (In the case of adding neutral comments) we found that Increasing the number of comments to the functionality of (test 1, test 3, test 7, test 6) improved the results of all classifiers.

Based on tests (4), (11), (2), (12), (8), (15), (5) and (16) we have observed in these tests that the balance and imbalance of the number of comments have the same effect weight.On the other hand, in tests (1), (9), (3), (10), (7), (13), (6) and (14), we noted that the balance and imbalance of the number of comments did not have the same Impact weight.

- ❖ We conclude that the improvement of the result is not due to imbalance but because of the largest of the dataset.

5.4 Environment and Development Tools:

For programming, we used the Python environment.

5.4.1 Python language:

Python is a multi-paradigm programming language. It promotes structured, object-oriented imperative programming. It has a strong dynamic typing. Developed by "Python Software Foundation". Its implementation started in December 1989.

It has several versions Python 2. x or Python 3. x and we will use our latest version of Python 3.7.3. [32].

5.4.1.1 What is python for?

Python offers high quality libraries, the most used in artificial intelligence and machine learning. It supports TensorFlow, Tweepy and Python Pyspark. In the field of artificial intelligence, we find that the most famous libraries are made of python, such as Scikit-Learn and Pandas. [33].

5.4.1.2 Scikit-Learn:

This is a free software machine-learning library for the Python programming language. It features various classification, regression, and grouping algorithms, including support vector machines, random forests, gradient accelerators, k-means, and DBSCAN, and is designed to interact with Python digital and scientific libraries NumPy [34].

5.4.1.3 Pandas:

Is an open source library under BSD license providing high performance and easy to use data structures, as well as data analysis tools for the Python programming language [35].

5.4.1.4 CSV:

library in t Python used for import or export databases of csv format files , and includes all the necessary functions : csv.reader , csv.writer ….. etc[34]

5.5 Configuration material

we used two PCs, the first:

- ✓ Processor: Intel® Core ™ i3-3217U CPU @ 1.80 GHz.

- ✓ RAM: 4 GB.
- ✓ System type: Ubuntu 16.04.6 LTS 64 bit operating system.

And the second:

- ✓ Processor: Intel® Core ™ i3-2377M CPU @ 1.50 GHz.
- ✓ RAM: 4 GB.
- ✓ System type: Windows 7 32 bit operating system.

5.6 Conclusion

In this chapter, we implemented four algorithms of machine learning. These algorithms are: Support Vector Machines (SVM), Tree Decision (DT), Random Forest (RF) and Naïve Bayes (NB)..on the datasets and display their results to obtain a modal. And we operated four machine learning classifiers that support vector machines (SVM), decision tree (DT), random forest (RF) and naive Bayes (NB) on a set of text containing 3373 text, described as follows: 701 Islamic text, 643 Christian text, 716 Jewish text ,706 Atheism text and 607 neutral text. where these classifiers are evaluated by 20% of the corpus .We used eighteen(18) Functionality that are:the first was conducted using four religious classes, and the second was performed on five classes (Four religious classes and one neutral class), The rest of the functionalities were performed on two classes with the balance or imbalance of text property.

We compared test results for the four classifiers. We found that the great accuracy (93.77%) is achieved by the Random Forest (RF) classifier and support vector machines (SVM) is generally regarded as a better classifier..

6. General Conclusion :

The objective of this thesis is revealing of peoples' religion in social networks with Five categories: (Muslim,Christian,Jewish, atheist, and other neutral) .The goal of our work is the realization of an application under Python which uses a data source (.csv) contains texts noted by values: (1,2,3,4,5). Also based on a lexicon of words (.csv) to classify these texts.

In this work, we have classified comments on a corpus, which contains 3373 religious texts Almost Classified as the following: 701 Muslim commentary, 643 Commentary on Christ and 716 Jewish commentary.706 commentary for Atheists and 607 neutral commentary.We used four machine learning classifiers which are support vector machine (SVM), decision tree (DT), forest of decision trees (RF), naive Bayesian (NB) where the evaluations of these classifiers is done. By 20% of the corpus.

We have used this feature in our work which is the Dictionaries of the studied religions which are The Quran , The Prophetic Hadith ,The names of the prophets and messengers ,The names of the companions ,The supplications ,The religious months, The names of religious scholars ,The Torah ,The Bible ,The names of the rabbis . We conducted 18 different tests, one using all categories (five classes) and the other using only religious categories (four classes). The rest of the tests were compared with two classes, and we compared the test results for the four classifiers.

In the first Chapter, we studied the different definitions that we need in our thesis. Where we provided a comprehensive definition of artificial intelligence and we defined machine learning and mentioned the most important methods that it uses, including supervised learning and unsupervised learning.

In the Second Chapter, we have defined the concept of social network analysis then cite some areas of use for this analysis. Subsequently we introduced the methods that allow the analysis of social networks.

After that, we presented the definition of text mining and NLP, we mentioned how to categorize a text and his mechanism.finally we defined Machine Learning for Natural Language Processing.

The Third Chapter is devoted to the study of the four researches.They generally talk about the ways in which religions are classified. Each research uses its own method.

Where we presented the ways and methods used in their work and presented the results that they reached, then we compared these results to show which methods were better.

In the fourth chapter, we presented the datasets and dictionaries that we used in our study, and we defined the features .

Finaly we implemented four algorithms of machine learning: Support Vector Machines (SVM), Tree Decision (DT), Random Forest (RF) and Naïve Bayes (NB) to obtain a modal. And We presented the results obtained and discussed them.

We find that the best algorithm is Random Forest (RF) with 93.77% and SVM with 93.33%. It should be noted that in our attempt to improve the results obtained, we have implemented classification algorithms by comparing only two classes instead of comparing them with all the classes .

Outlook:

Our work is in the theme of Doctorat with the teacher MEFTAH and the Mr ADEL that is about discovering extremism.

Our work will be as an intial phase after precising religious categories.

The discovery of extremist text will be with each classifier from religious classification.

Morover the reached results of our study can be applied on other languages such as Arabic.

Bibliography:

[1] MANSOUR Abd El Fattah, TRAD Aissa, Automatic detection of abusive speech, offensive and obscene in the Algerian dialect, End of Study Thesis Presented for obtaining the ACADEMIC MASTER Diploma, ECHAHID HAMMA LAKHDAR UNIVERSITY OF EL OUED.

[2]Jim Goodnight and Oliver Schabenberger,and Simon Sinek,site web,sas ,2020 ,what-is-artificial-intelligence,https://www.sas.com/en_us/insights/analytics/what-is-artificial-intelligence.html

[3] 7wData, August 9, 2018, https://www.7wdata.be/business-analytics/artificial-intelligence-what-it-is-and-why-it-matters/

[4]builtin, 2019,https://builtin.com/artificial-intelligence

[5] Lina Klass, spotlightmetal,this article was first published on MM International, ID: 45469648, 29.08.2018,https://www.spotlightmetal.com/machine-learning--definition-and-application-examples-a-746226/?cmp=go-aw-art-trf-SLM_DSA-20180820&gclid=Cj0KCQiA0NfvBRCVARIsAO4930ngxcUMYgPcqOnWvEjCizZtc83wefDtyydlCLCf-BiCmM7NXB-dzs0aAr4WEALw_wcB

[6] Buzz Blog Boxhttps,August 03/ 2019,://becominghuman.ai/what-is-training-data-its-types-and-why-it-is-important-f998424c3c9

[7] Shubham Patni,Know about the types of Machine Learning, May 7, 2018,https://medium.com/@shubhapatnim86/know-about-types-of-machine-language-part-2-19ca69071ed

[8] Software Testing Help, Tutorial, Last Updated April 16, 2020 https://www.softwaretestinghelp.com/types-of-machine-learning-supervised-unsupervised/

[9] Krishna,guru99,tutorial ,2020, https://www.guru99.com/unsupervised-machine-learning.html

[10] Mohammad Waseem, edureka,Published on Dec 04/2019 ,https://www.edureka.co/blog/classification-in-machine-learning/

[11] Subrat,Tutorial , Simply learn ,https://www.simplilearn.com/classification-machine-learning-tutorial

[12]BENSALEM Khadidja SOUALMI Samiha, Detection of rumors in social networks, Master thesis, Abderrahmane Mira University –Bejaia, 2016/2017

[13] Djellouli Mohamed and TABTI Abdelkrim ,MEMOIRE DE Master, Analyse De Contexte Des Personnes Sur Les Reseaux Sociaux : Analyse Des Sentiments Sur Twitter ,université DR.taher moulay SAIDA faculté Technologie déparement :Informatique ,2017 – 2018 .

[14] Pew research center,soccial media update 09/01/2015, https://brgcommunications.com/social-media-usage-statistics-2015/

[15] Eve Michael. “Two traditions of social network analysis”, Networks no 115. L 183-212. 2002

[16] Mercanti-Onérin Maria. Analysis of social networks and online communities: What applications and marketing. Center (research D \ ISP, DRM (CNRS UMR 7088), University Paris Dauphine. FRANCE. 2010.

[17] Maria Malek. Laris-Eisti. Introduction to social network analysis. 2009.

[18] Erick Stattner. Introduction to Social Network Analysis. LAMIA Laboratory University of the Antilles and Guyana. Guadeloupe, FRANCE. 2012

[19] Matthew. Gaëtan. Benedict. Stepliane. Search of online social networks. Bibliographic summary on Social Data Mining. 2012

[20]allbrandworks, Article,6 November2019 ,http://allbrandworks.com/2019/11/06/the-impact-of-artificial-intelligence-on-social-media/

[21] David Milward,website Linguamatics,2019 https://www.linguamatics.com/what-text-mining-text-analytics-and-natural language processing?fbclid=IwAR3evoySEwRw0hPU3gLOEYLlhGR3how7c1yipzr3CyBv8Nas0TEGpSOKcPA

[22] Mr. Aouine Mohammed, Automatic categorization of arabic text, memory of magister for the graduation of presented at the department of data processing magister in data processing option artificial intelligence and imaging university of guelma faculty of sciences and engineer, 2009.

[23] BENAISSA Bedr-Eddine, Magister's thesis ’semi-automatic construction of ontologies from Arabic texts. Abou Bekr Belkaid Tlemcen University Algeria. 2011/2012.

[24] Lamamri Nadia, Thesis Presented for the Graduation of Computer Science Master, Analysis of Opinions Expressed in the Form of Arabic Texts, Laarbi Ben M’hidi University - OEB, 2013/3014

[25] Abderrahim, MA: A morphological analyzer for vowel or not Arabic. SIIE'2008: 1st International Conference, Information Systems and Economic Intelligence, SIIE 2008 Hammamet - Tunisia, February 14-16, 2008, Proceedings tome II, IHE éditions, ISBN 9978-9973-868-20-6, pp. 324--339 (2008).

[26] Mohammed El Amine Abderrahime 2nd international conference on computing and its applications. Towards an interface for enriching queries in Arabic in an information retrieval system. Abou Bekr Belkaid Tlemcen University Algeria. 2009.

[27] Lexalytics, Paul Barba, Machine Learning for Natural Language Processing, November 25, 2019,https://www.lexalytics.com/lexablog/machine-learning-natural-language-processing?fbclid=IwAR0iRYxEjE_Rke597GVwli0-B2oa0BPLWpVNdLalgclZV9Hr32dIDvzZnfQ

[28] Minh-Thap Nguyen,On Predicting Religion Labels in Microblogging Networks, Living Analytics Research Centre Singapore Management University,2012.

[29] Shakeel Ahmad, Muhammad Zubair Asghar,Fahad M. Alotaibi And Irfanullah Awan, Detection And Classification Of Social Media‑Based Extremist Affiliations Using Sentiment Analysis Techniques, Research,2019 .

[30] Mayuri Verma,Delhi,Technological University, Lexical Analysis of Religious Texts using Text Mining and Machine Learning Tools, June 2017.

[31] Sanja ˇStajner, Ruslan Mitkov, Style of Religious Texts in 20th Century,Research Institute in Information and Language Processing University of Wolverhampton, UK,

[32]amazon.fr,15février2018, https://www.amazon.fr/gp/product/241203446X/ref=as_li_qf_asin_il_tl?ie=UTF8&tag=pepsmultime0a-

21&creative=6746&linkCode=as2&creativeASIN=241203446X&linkId=660e1082f5
8229153d3457241e5db28b#reader_B07HHM72D1

[33]Pythonforbeginners,January18,2013,
https://www.pythonforbeginners.com/csv/using-the-csv-module-in-python

[34] scikit-learn, https://scikit-learn.org/

[35] Pandas, May 28, 2020, https://pandas.pydata.org/