T.C

ALTINBAS UNIVERSITY

Graduate School of Science and Engineering

Information Technologies

AUTHORSHIP ATTRIBUTION FOR SOCIAL MEDIA

Essam Amer Owen Shanna

M.Sc. Thesis

Supervised by Prof. Dr. Oğuz Bayat

Istanbul,2019

AUTHORSHIP ATTRIBUTION FOR SOCIAL MEDIA

By

Essam Amer Owen Shanna

Information Technologies

Submitted to the Graduate Faculty of

Altinbas Universities in partial fulfillment

Of the requirements for the degree of

Master of Information Technologies.

ALTINBAS UNIVERSITY

2019

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of .....

Academic Title Name SURNAME	Academic Title Name SURNAME
Co-Supervisor	Supervisor

Examining Committee Members (first name belongs to the chairperson of the jury and the second name belongs to supervisor)

Academic Title Name SURNAME	Faculty, University	_____
Academic Title Name SURNAME	Faculty, University	_____
Academic Title Name SURNAME	Faculty, University	_____
Academic Title Name SURNAME	Faculty, University	_____
Academic Title Name SURNAME	Faculty, University	_____

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

Academic Title Name SURNAME

Head of Department

Academic Title Name SURNAME

Director

Approval Date of Graduate School of Science and Engineering: ____/____/____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Essam Amer Owen Shanna

ACKNOWLEDGEMENTS

I'm deeply grateful to my supervisor Prof. Dr. Oğuz Bayat for his guidance, his time, and invaluable comments on the thesis.

I would like to express my deepest appreciation to my friends for their unconditional support.

Also, I would thank my country Libya for the unlimited support.

My final words go to my family. I want to thank my family, whose love and guidance is with me in whatever I pursue.

ÖZET

SOSYAL MEDYADA YAZARLIK ATIF

Essam Amer Owen Shanna

Yüksek Lisans Bilgisayar Mühendisliği Bölümü, Altınbaş Üniversitesi

Danışman: Prof.Dr. Oğuz Bayat

Tarih: Ağustos, 2019

Sayfalar: 61

Tarihsel olarak Yazarlık Atıf Teknikleri 1887'de ortaya çıktı, Mendenhall ilk kez yazarın kişiliğini belirtmek için Özellikler Sayma fikrini icat ettiğinde [1]. Metin için Yazarlık Özelliği, bir metnin yazarının diğer eserlerini inceleyerek benzerliğini belirlemeye odaklanır.Yazı koleksiyonunda ve yazı stilinin yakalanmasında nispeten sabit kalan bir dizi metin metnin özniteliklerini ortaya koyması durumunda, Bu araştırma, bir yazar tarafından yazılan tweet'lerden bir grup özelliğin çıkarılmasını ve bu özellikleri makineye besleyerek, yazarın yazar grubundan tweet yazıcısını tahmin edebilmek için yazar yazım tarzını öğrenmesini sağlar.

Anahtar kelimeler: Yazar Tanıma, sosyal medya, makine öğrenmesi, NLP, özelliklerin çıkarılması, Twitter

ABSTRACT

AUTHORSHIP ATTRIBUTION FOR SOCIAL MEDIA

Essam Amer Owen Shanna

M.Sc. Information Technologies, Altinbas University,

Supervisor: Prof. Dr. Oğuz Bayat

Date: August 2019

Pages: 61

Historically the Authorship Attribution Techniques appeared in 1887 when Mendenhall first invented the idea of Counting Features to indicate the personality of the author [1]. The Authorship Attribution for the text author focuses on identifying the similarity of the author of a text by examining his other works. Where a set of features of text extracted which that remain relatively constant in his collection of writings and the capture of his writing style, this research discussing extracting a group of features from tweets written by an author and feed this features to the machine to learn the author writing style to be able to predict the tweet' author from authors group.

Keywords: Authorship attribution, social media, machine learning, NLP, extracting features, Twitter

TABLE OF CONTENTS

Pages

LIST OF TABLES

Pages

LIST OF FIGURES

Pages

1. INTRODUCTION

Historically the Authorship Attribution Techniques appeared in 1887 when Mendenhall first invented the idea of Counting Features to indicate the personality of the author [1] This was followed by the work of: Yule (1938) and Morton (1965) using sentence lengths to judge the author's identity [2]. Then the studies continued in this field in many languages.

The Authorship Attribution for the text author focuses on identifying the similarity of the author of a text by examining his other works. Where a set of features of text extracted which that remain relatively constant in his collection of writings and the capture of his writing style; in that way we can represent the author by N dimensions where N is the number of characteristics extracted from the text [3].

The process of attribution is centered on studying the characteristics of the text in order to draw conclusions about its composition, it originated mainly from the science of stylometry which is a branch from the linguistics science which applies statistical measurements to the literary method. The Authorship Attribution can be used in a wide range of applications, Anonymous or Conflicting Books, The disclosure of plagiarism in order to indicate whether the author is the real author, is also the legal investigation to ascertain the authors of e-mails and newsgroups [4]. In recent years, these applications have grown in various areas such as intelligence, criminal law, civil law. It's called Forensic authorship attribution which "assumed that the source of a text is either one author (or possibly several) out of a known set, or an author that is unknown to investigators" [5]

Brocardo et al. in their research defined three types of authorship analysis: authorship attribution which about "determining the most likely author of a target document among a list of known individuals". [6]

The statistical work carried out by the intensive work of the experts at the first time has become computer-aided, and in addition to the use of the high processing power of the computers in statistical processes, machine learning techniques have also begun to be implemented. Where the works that make use of computers and on the electronic documents are represent the "modern" and the traditional one is done on a handwritten document and by expert labors.

Nowadays and with the spreading of the social media in our life, the needs for identify the author is appeared again, where there is a need to know the writer of a tweet or post between many writers maybe to discover a threat or detect a person tweets from a fake account under a fake identity.

But the social media platforms have many different characteristics, and it's not like a literature works, in the social media we have a limited size of letters to express our ideas, there are no specific rules to write, there is no grammars restrict; in the same time it has other features which represent the post identity like the date, location …

In this study we will discuss the ability to make the authorship attribution on the social media by extracting a feature from the posts, this feature will represent the "writer print".

1.1 LITERATURE REVIEW

In this section, a brief summary for the literature which discussed the Authorship Attribution, properties and methods studied:

Castro and Lindauer in their research on authorship attribution on twitter data, they used dataset from 800 users each one has 1000 tweets at minimum extracted by using Twitter API , the authors used some website to know the twitter users who have two accounts at the same time such: Dell Solicitation; also used the web search by searching in Google about phrases like "also follow me at on Twitter profile pages", beside scrapping some Google + profiles who their owners mentioned that they have many twitter accounts . And due to the small number of retrieved accounts which mate the required conditions, the authors simulate the dual authorships through splitting each feed. The authors extracted 393 features using Ruby script, the extracted features stored in CSV file to transform to the classifier which implemented in MATLAB tool. As table (1.1) shows [7]:

Table 1.1: Extracted features [7]

Category	Description	Count
Length	Words/characters per post	2
Word shanpe	Frequency of words in uppercase, lowercase, capitalized, camelcase, and other capitalization schemes	5
Word length	Histogram of word lengths from 1-20	20
Character frequencies	Frequency of letters a-z (ignoring case), digits, and many ASCII sym-bols	68
Unicode	Frequency of non-ASCII characters	1
Function/stop words	Frequency of words like “the”, “of”, and “then”	293
Twitter conventions	Existence or “RT” or “MT”	2
Retweets	Whether the post is an exact retweet or a modified retweet	2

The authors used the information gain algorithm to rank the features set to discover the feature set which give the higher prediction accuracy. The authors used a combination of nearest neighbor’s algorithm (NN) and regularized least squares classification (RLSC), the used algorithm classified 41% of the twitter known dual accounts correctly [7].

Rabab’ah et al. in their research made authorship attribution on Arabic tweets extracted by Twitter API, through fetching 38,386 tweets for 12 twitter accounts from the top 100 Arab influencers in twitter writing in different domains, who have the largest followers numbers; the authors cleaned the tweets to remain 37,445 tweets. Then extracted 951 features related with linguistic features including: Part Of Speech (POS), which according to the authors didn’t used before in the Authorship Attribution researches, the stylometric features, and Bag-Of-Words (BOW) features. The contribution of the research according to the authors is combining all the mentioned features sets and learn the machine on them; where after applying many classifiers on the datasets: Naïve Bias (NB), decision tree (DT), Support Vector Machine (SVM) the result for each feature set and for the combination of all the features sets were as table (1.2) shows [8]:

Table 1.2: Features combination [8]

Feature Set	Number of Features	NB	DT	SVM
MF	57	24.73%	28.61%	28.53%
ASF	340	31.22%	49.65%	55.77%
BOW	554	46.55%	53.02%	57.18%
MF+BOW	611	42.57%	52.06%	59.19%
MF+ASF	397	32.32%	51.55%	57.51%
All	951	38.35%	59.83%	68.67%

Rao et al. in their research make authorship attribution on 200 news articles token from four authors, each one has 50 articles, the authors represents the corpus as word N-gram where n= 1,2,3; then they calculated the Term Frequency and Inverse Document Frequency (TF-IDF) to the documents to represent them in Vector Space Model. The K-means, Mini Batch K-means and Ward Hierarchical Clustering algorithms have been used as unsupervised classification algorithms [9]. The accuracy prediction for each author was as shown in table (1.3) shows [9]:

Table 1.3: Accuracy prediction for each author [9]

	% of Cluster Rate		
Author Name	K-means	Mini Batch K-means	War Hierarchical
Aaron Pressman	100	68	64
Benjamin Kang Lim	100	96	100
Darren Schuettler	96	100	100
David Lawder	92	78	76

Where the average prediction accuracy for all the dataset according to each clustering algorithm was as shown in the table (1.4) shows [9]:

Table 1.4: Average prediction accuracy [9]

Clustering algorithms	K-means	Mini Batch K-means	Ward Hierarchical
	97	85	85

Rocha et al. in their research made authorship attribution on dataset contains 50000 tweets, wrote by 50 users for each one 1000 tweets; the dataset cleaned by removing all non-English tweets, less than 4 words tweets, retweets or tweets that contains RT tag; replacing numbers, URLs,

dates and timestamps by the meta tags NUM, URL, DAT, and TIM. Then the features extracted from the tweets as following: function words, TF-IDF, the character and word n-gram, The availability of natural language processing (NLP) toolkits, Part-of-Speech (POS) Tagging, and POS trained for twitter to extract the retweet and links, special POS to extract hashtag, mention, URLs and emoticons; the authors applied Mean SVM , W-SVM , Random Forests algorithms on the data set, the figure (1.1) shows the accuracy prediction for Mean SVM classifier, the figure (1.2) shows the accuracy prediction for W-SVM classifier, the figure (1.3) shows the accuracy prediction for Random Forests classifiers; for each user group [10]:

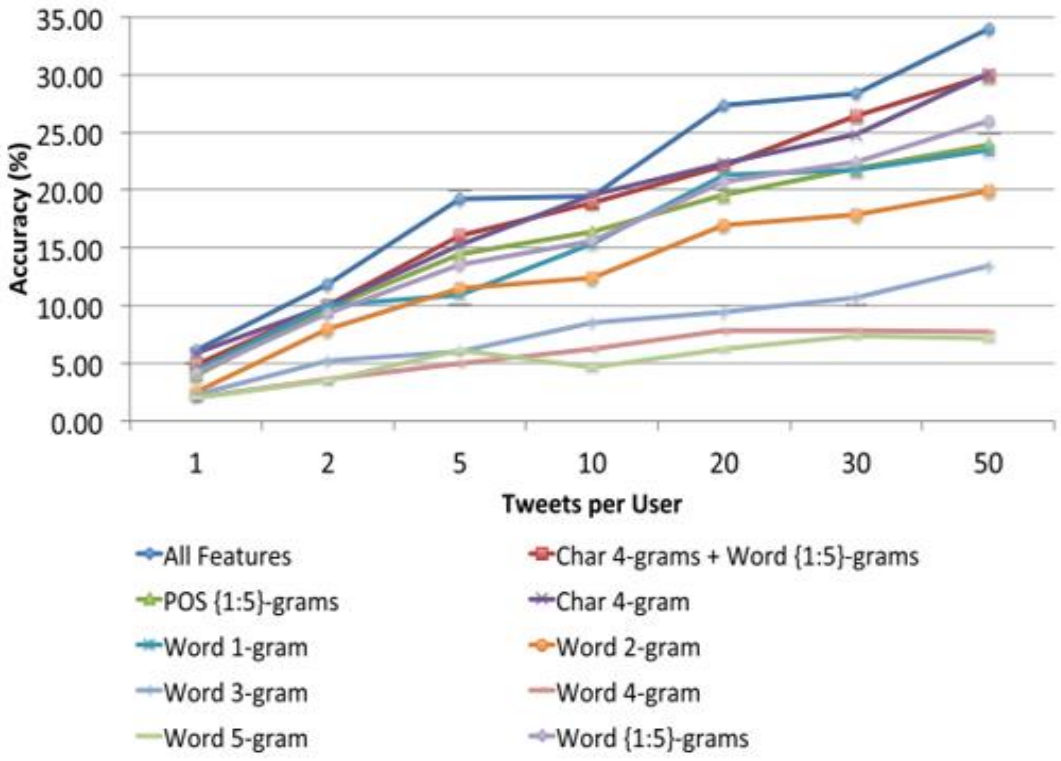

Figure 1.1: The accuracy prediction for Mean SVM classifier [10]

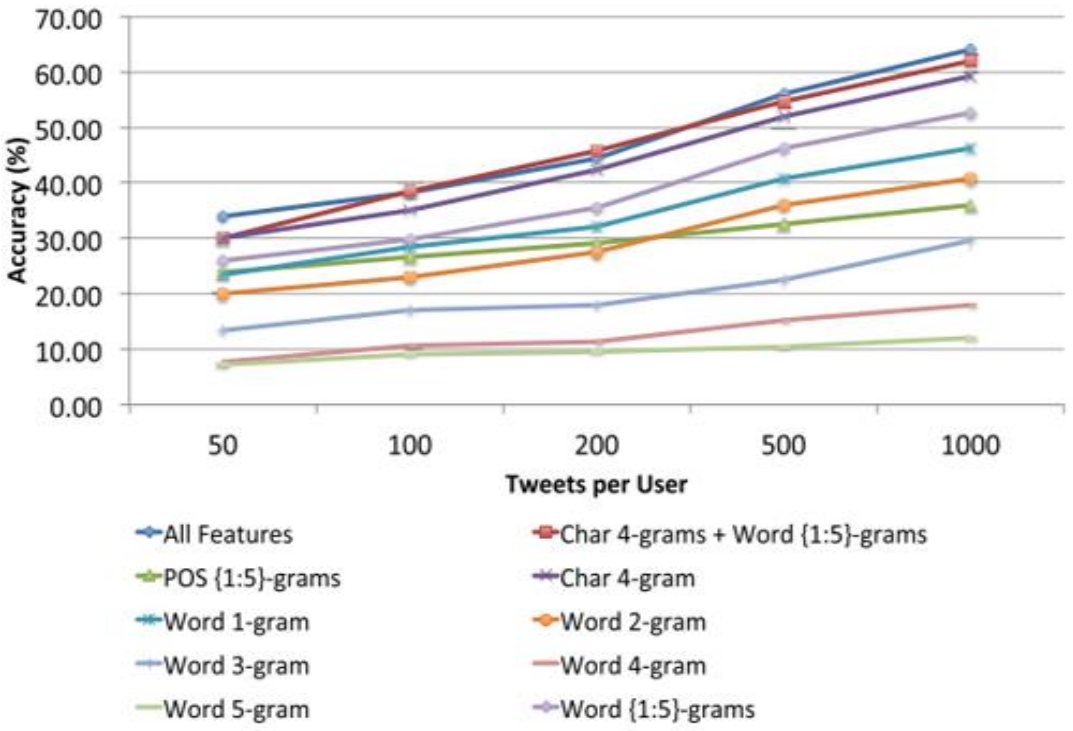

Figure 1.2: The accuracy prediction for W-SVM classifier [10]

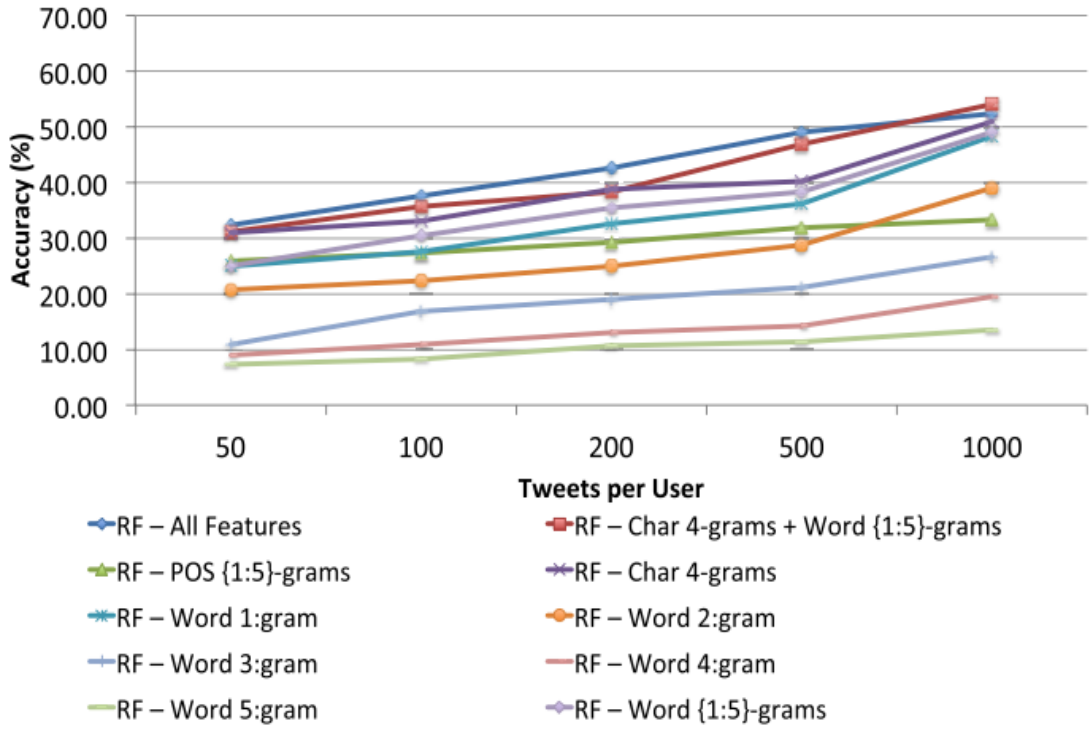

Figure 1.3: The accuracy prediction for Random Forests classifiers [10]

Usha et al. in their research made authorship attribution on 200 users with 800 tweets, the authors labeled the dataset on 5 personality traits to train the model; the dataset tested manually, where the very small tweets combined together expressing one tweet; the authors used the personality, LIWC (Linguistic Inquiry Word Count), psycholinguistic and tone based features (anger, sad, disgust, fear, surprise, joy humor and sarcasm). The proposed model from in this research was as figure (1.4) shows [11]:

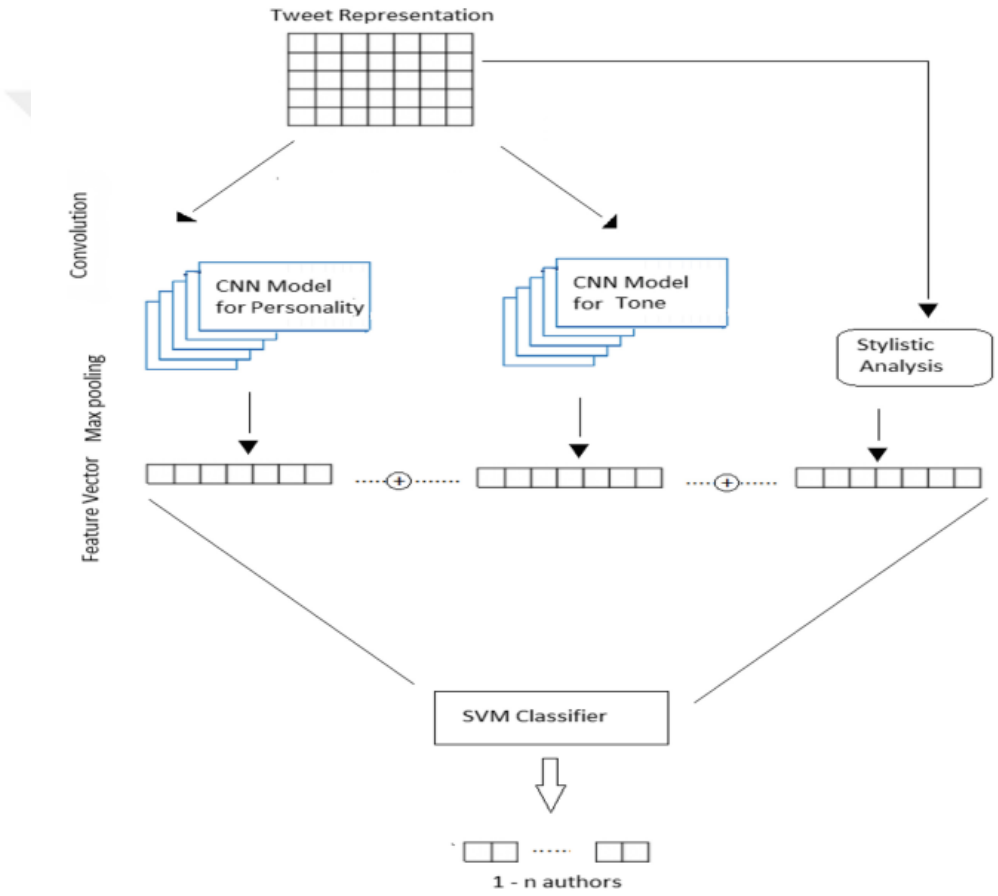

Figure 1.4: USHA proposed model [11]

The above model contains many steps:

- Using the transfer base Conventional Neural network (CNN) model to identify the user tone
- Using CNN to identify the personality

- Employ a combination between the stylistic features and features extracted from the CNN models , then will be import to SVM classifier

According to the author the above model give better prediction accuracy with more number of candidate authors, the table (1.5) shows the accuracy for each user groups by using tow tweet data set for each author 250 or 800 tweets [11]:

Table 1.5: The accuracy for each user groups by using tow tweet data set [11]

User	Tweets per user	precision	Recall	F-score	Accuracy
15	250	0.55	0.50	0.52	51%
	800	0.85	0.79	0.81	80%
50	250	0.50	0.48	0.49	50%
	800	0.77	0.69	0.70	71%
100	250	0.44	0.49	0.46	47%
	800	0.66	0.52	0.58	58%
200	250	0.39	0.31	0.34	38%
	800	0.59	0.63	0.54	53%

In our work we will continue this effort, by adding more features such semantic features, and nonlinguistic features, which according to our knowledge were not used before.

1.2 PROBLEM DEFINITION AND PROPOSED SOLUTION

Author Attribution is the problem of finding an author's written knowledge within a possible set of authors by examining an unknown text. At first glance, it can also be seen as a problem of classifying texts according to authors. The features used in classification consist of various statistical values specific to the lexical, syntactic, structural, content, or article extracted from texts. Thousands of features have been used in many studies over time [7], but a set of features that have been widely accepted and agreed upon have not been identified.

Machine learning techniques are used frequently and successfully in the definition of author as well as in many classification problems. The feature vectors used in the training of predictive models generated in supervised learning studies are defined by the features extracted from the texts. Although many features can be extracted from the texts, the predictor model of the feature vectors contains a large number of elements, which complicates the processing time required for

training and classification. On the other hand, the use of more textual features does not mean that authors can be identified with higher percentages.[6] The small feature vectors that can simply be extracted from texts are important for the development of systems that can be quickly and easily adapted and that do not require a lot of processing power. Different languages may have different characteristics. Therefore, the presence of feature clusters of language-specific and minimal features that can yield successful classification results is seen as a significant problem.

Also in the social media there are a new cluster of features which is not related with the textual features, and this features extracted from the platform itself not the text, where each platform gives a specific features to the writer or post to define it, and these features different from a platform to another according to the platform policy.

In this study we suppose that if we combine a textual features with the no textual features on a data set extracted from the social media for a set of authors and a machine learning model built to classify the author's writings according to the authors, we can recognize the author of unrecognized post.

2. THEORITICAL FRAMEWORK

2.1 RESEARCH FEASIBILTIY

Actually there are many accidents happened due to using fake profiles, hacking other profiles , or imitating their style in writing, and these accidents have many dangerous levels according to the used social media platform and the hacker intentions against the victims, whereas some social media platform are allowing to use username similar to other users like Myspace where in one case a mother of one of the teenage girls made a fake account on Myspace similar to the account of her daughter's lover, she put his picture with the personal information, and started chatting her girl; at the end of the messaging she sent a message imitating her lover style informing the girl that he no longer like her and the life is beautiful without her, the girl committed suicide after a wave of nervous breakdown, The mother immediately wiped the account . [13]

A young American girl after she left the job disappeared for a short time, and since her absence was normal, her family did not notice her absence, but someone made Facebook account using the girl information for the purpose of frightening her family, the first message was that the girl was killed. The family began to send messages to the people of the neighborhood to search for the girl in the neighborhood, but he began to send messages that he was watching the neighborhood and asked them to stop searching. Fortunately, the girl's story spread in local media and investigated the matter, which lead to arrest of six people involved in the Kidnapping and murder operation. [13]

as example for social media accounts hacking and the exploitation for trust of a user's network of friends, it’s good to mention the Bryan Rutberg story a Microsoft employee, where a hacker intrusion his Facebook account, and sent a message to Brian's friends telling them he was in financial trouble and that they had to immediately transfer money through the Western Union to London, The letter contained information on how to send money to him. At the same time, the hacker changed the account settings to prevent Brian from login his account and unfriend his friendship with his wife to prevent her from telling friends that the account was hacked [13].

In general, the accounts of Social Media have become important in investigations and are considered evidence in any issue. From here it good to search for a model of method to authenticate the author from his writing style, and try to extract features from his wrote being like his fingerprint

in writing, which may support or refute any claim when using the Social Media account as evidence in any charge with legal or moral consequences

And that what be handled in this research, where a method to discover the author of tweet between many authors will be developed using textual and non-textual features extracted from tweets to learn the machine algorithms on them to be able to predict the tweet author by feeding with a tweet from Anonymous or dubious source.

2.2 CLASSIFICATION

In general, the classification process is to assign individual to one class between many different classes, but there are many classification types:

2.2.1 Binary Classification

Binary classification problems [13], consider assigning an individual to one of two categories, by measuring a series of attributes. An example is authorship for a single tweet or document assign to which one of two authors belong according the author writing style. It's not appropriate for this study.

2.2.2 Multiclass Classification

Consider assigning an individual to one of more than two categories [14]; as in the research case where for the authorship attribution problem the goal is assigning a tweet to one of group of supposed authors, according to the author's style in the writing.

2.3 MACHINE LEARNING

2.3.1 History

The history of AI started in the year 1943 when Waren McCulloch and Walter Pitts introduced the first neural network model. Alan Turing introduced the next noticeable work in the development of the AI in 1950 when he asked his famous question: can machines think? He introduced the B-type neural networks and also the concept of test of intelligence. In 1955, Oliver Selfridge proposed the use of computers for pattern recognition.

In 1956, John McCarthy, Marvin Minsky, Nathan Rochester of IBM, and Claude Shannon organized the first summer AI conference at Dartmouth College, the United States. In the second Dartmouth conference, the term artificial intelligence was used for the first time. The term cognitive science originated in 1956, during a symposium in information science at the MIT, the United States. Rosenblatt invented the first perceptron in 1957. Then in 1959, John McCarthy invented the LISP programming language.

David Hubel and Torsten Wiesel proposed the use of neural networks for the computer vision in 1962. Joseph Weizenbaum developed the first expert system Eliza that could diagnose a disease from its symptoms.

The National Research Council (NRC) of the United States founded the Automatic Language Processing Advisory Committee (ALPAC) in 1964 to advance the research in the natural language processing. But after many years, the two organizations terminated the research because of the high expenses and low progress. Marvin Minsky and Seymour Papert published their book Perceptron in 1969, in which they demonstrated the limitations of neural networks.

As a result, organizations stopped funding research on neural networks. The period from 1969 to 1979 witnessed a growth in the research of knowledge-based systems. The developed programs Dendral and Mycin are examples of this research. In 1979, Paul Werbos proposed the first efficient neural network model with back propagation. However, in 1986, David Rumelhart, Geoffrey Hinton, and Ronald Williams discovered a method that allowed a network to learn to discriminate between nonlinear separable classes, and they named it back propagation.

In 1987, Terrence Sejnowski and Charles Rosenberg developed an artificial neural network NET Talk for speech recognition.

In 1987, John H. Holland and Arthur W. Burks invented an adapted computing system that is capable of learning. In fact, the development of the theory and application of genetic algorithms was inspired by the book Adaptation in Neural and Artificial Systems, written by Holland in 1975. In 1989, Dean Pomerleau proposed ALVINN (autonomous land vehicle in a neural network), which was a three-layer neural network designed for the task of the road following.

In the year 1997, the Deep Blue chess machine, designed by IBM, defeated Garry Kasparov, the world chess champion. In 2011, Watson, a computer developed by IBM, defeated Brad Rutter and Ken Jennings, the champions of the television game show Jeopardy!

The period from 1997 to the present witnessed rapid developments in reinforcement learning, natural language processing, emotional understanding, computer vision, and computer hearing.

The current research in machine learning focuses on computer vision, hearing, natural languages processing, image processing and pattern recognition, cognitive computing, knowledge representation, and so on. These research trends aim to provide machines with the abilities of gathering data through senses similar to the human senses and then processing the gathered data by using the computational intelligence tools and machine learning methods to conduct predictions and making decisions at the same level as humans.

The term machine learning means to enable machines to learn without programming them explicitly [15].

2.3.2 Definition

Machines are by nature not intelligent. Initially, machines were designed to perform specific tasks, such as running on the railway, controlling the traffic flow, digging deep holes, traveling into the space, and shooting at moving objects [18].

Machines do their tasks much faster with a higher level of precision compared to humans. They have made our lives easy and smooth. The fundamental difference between humans and machines in performing their work is intelligence. The human brain receives data gathered by the five senses: vision, hearing, smell, taste, and tactility. These gathered data are sent to the human brain via the neural system for perception and taking action. In the perception process, the data is organized, recognized by comparing it to previous experiences that were stored in the memory, and interpreted. Accordingly, the brain takes the decision and directs the body parts to react against that action. At the end of the experience, it might be stored in the memory for future benefits [16].

A machine cannot deal with the gathered data in an intelligent way. It does not have the ability to analyze data for classification, benefit from previous experiences, and store the new experiences to the memory units; that is, machines do not learn from experience.

Although machines are expected to do mechanical jobs much faster than humans, it is not expected from a machine to: understand the play Romeo and Juliet, jump over a hole in the street, form friendships, interact with other machines through a common language, recognize dangers and the ways to avoid them, decide about a disease from its symptoms and laboratory tests, recognize the face of the criminal, and so on. The challenge is to make dumb machines learn to cope correctly with such situations. Because machines have been originally created to help humans in their daily lives, it is necessary for the machines to think, understand to solve problems, and take suitable decisions akin to humans. In other words, we need smart machines. In fact, the term smart machine is symbolic to machine learning success stories and its future targets. The question of whether a machine can think was first asked by the British mathematician Alan Turing in 1955, which was the start of the artificial intelligence history [16].

Whereas In traditional programming, the programmer specifically informs the program what to do in each case and should think about all the possibilities that can happen to write what is appropriate to act with in the program if it occurs. This method of programming is very common and is used for programming applications of mobile phones, websites, and computer programs. One of the problems with this method is that sometimes you encounter applications that are difficult to limit to all the possibilities that will be activated when you run them. Take, for example, a program to identify people by the face. How will you tell the program about facial recognition? Consider the enormous differences and contrasts between the faces. as shown in figure (2.1) shows [20].

Figure 2.1: Machine learning on image processing [20]

In Machine Learning algorithms, the program learns what to do based on data automatically, without specific instructions from the programmer. In the Face Recognition example, the program automatically extracts the distinguishing features that help it to differentiate between different faces, and then uses it when you insert a new face image to recognize it automatically. The process of drawing distinctive features is at the learning or training stage, so the program can be used and confirmed in the test phase when a new image is inserted.

Enable the program to learn automatically Open vast horizons and applications, it was not possible or easy to work by directly programming. Some of these applications are used daily, such as speech recognition, text recognition in images and conversion to written, automated recommendation systems (such as in Amazon), search engines, recognition when there is a face in the picture (such as in cameras), marketing and advertising. There are also medical applications such as disease diagnosis, military, security, commercial, logistics (such as fleet deployment), and others.

There are so many machine learning algorithms, different schools follow in design and thinking. Some of them are based on statistical concepts such as probability and others using different theories such as graph theory, as well as a set based on heuristic rules and other design methods.

The common factor of all these algorithms is their attempt to find the best model, so that the data given is reduced in a way that ensures generalization when new data is used.

The following example explains the idea: Suppose we want to design a program that can tell if there are cats in a picture as shown in figure (2.2) shows [29]. Is it enough to bring several pictures of cats and teach the program to success in identifying new images? No. There are many forms and types of cats, is it enough to provide one image for each type and color known to cats? To take one of these pictures, let's say that for a cat in the middle of the picture, if we move the face up or right, can the program recognize it? Well, what if we take a close-up or distant shot of her whole body? Or if part of the face is covered? Or if the face is upside down? … Etc. We need an infinite number of images to cover it. The aim of generalization is to arrive at a model that can extract characteristics that indicate the presence of cats in general so that it can apply to any image, even if it is new, that is the essence of learning the machine.

Figure 2.2: Cat image from catadoptionteam.org website [29]

2.3.3 Machine Learning Techniques

There are four general machine learning methods: supervised, unsupervised, semi-supervised, and reinforcement learning methods.

The objectives of machine learning are to enable machines to make predictions, perform clustering, extract association rules, or make decisions from a given dataset, the following figure (2.3) briefly express the four methods and its specifications [17]

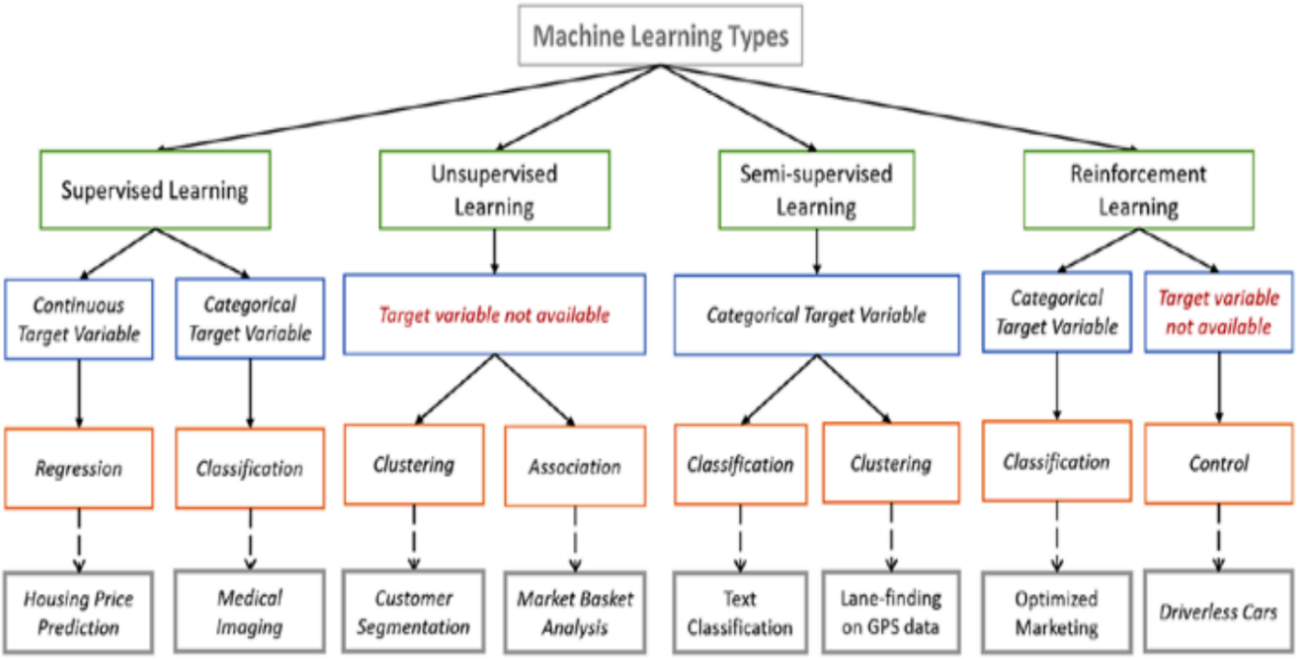

Figure 2.3: Machine learning types [17]

2.3.3.1 Supervised learning

In this type of classification, the system has a large set of documents and the system knows in advance the documents to any subsequent classification belong. These data are called training data, this data helps to train the classification function to know the amount of bias in each classification process. Most of the neural network algorithm classifiers follow this the behavior [18].

Two groups or categories of algorithms come under the umbrella of supervised learning. They are Regression and Classification.

Classification Algorithms: Learn something with specific classifications, such as colors, types of animals and fruit types. For example, classifying email as annoying or not.

Regression algorithms: The goal is to predict an undetermined number in a given category. Such as predicting temperature based on historical data, in addition to wind direction and clouds. As well as house prices based on data such as house size, location of the neighborhood, number of rooms and toilets.

2.3.3.2 Unsupervised learning

In unsupervised learning, there is no supervising on the data, or the data is unlabeled, due to many reasons, where in general labeling the data done manually or its need experts or for many reasons, where the goal is to find the hidden patterns inside the data [18].

In which the program combines similar data into groups, and then the test sample is classified based on proximity or further from these groups. One of the most famous algorithms in this category (K-Means). Applications can group customers with similar preferences in the same groups. Or the discovery and classification of communities (people of common interest) in social media.

Clustering is the most famous unsupervised learning algorithm, which in these the machine divide the instances into many clusters, each one contains the common instances [18].

Association: it's also important algorithm in the unsupervised learning, where in it the machine fine about the roles which make the instanced associated together [18].

2.3.3.3 Semi-Supervised learning

In this learning type, the used data are combination between the unclassified and classified data. The mixing of the unlabeled and labeled data use in generating appropriate model for data classification [18].

The semi supervised learning use to tech the model to be able to predict the test data better than the prediction done by just only the labeled data.

The semi supervised learning is imitating a baby behavior, where it's feed by two data sources:

- The data which feed by the environment as unlabeled, because its surroundings the baby
- The data which feed by the parents, for example the baby parents teach the baby about the objects names to distinguish them later, which is labeling process.

2.3.3.4 Reinforcement learning

The goal of this learning is to use the reward if the model learned correctly and the punishment if it's not, where the model is learning by the surrounding interactions [18].

2.4 PYTHON

Python is an interpreted programming language, object-oriented and dynamic writing. Python interpreter reads one line of code at a time, translates it into a low-level language (byte code) and then runs it. As a result, run-time errors are typically encountered. In fact, Python is one of the most popular programming languages because it is easy to program and understand. It is also open source. Created in 1991, by a developer named Guido Van Rossum.[19]

2.4.1 Python and Machine Learning

Machine learning is based on algorithms that automatically detect a specific pattern in the data being entered. Let's give an example. Suppose we want the learning program to determine whether the picture given to it contains a dog or a table. In this case, we first give the automated learning program 1000 pictures containing a dog and 1000 other pictures containing a table. The program will then learn the difference between the dog and the table. If you give it a new image, it will be able to recognize its content. This process is very similar to the way a child learns new things. It is very important not to tell the child explicitly that if there is something furry and has light brown hair, it is probably a dog. Maybe we just say: "This is a dog. This is also a dog. This is a table. This is also a table. " . In fact, machine learning algorithms work the same way. This applies to the recommendation system made by YouTube, Amazon and Netflix, as well as for facial or audio recognition programs [21].

Python is used in machine learning programs, there are many libraries and programming packages for machine learning in Python. Scikit-learn and TensorFlow are the best and most famous of these

libraries; scikit-learn is considering good for the beginners in machine learning, and TensorFlow considering good for the professional:

- Scikit-learn includes the most common machine learning algorithms.
- TensorFlow is a low-level library that allows the creation of custom machine learning algorithms.

According to the IBM report for the most important programming language for machine learning and artificial intelligence, python got the first rate as seen in the following figure (2.4) shows [21]

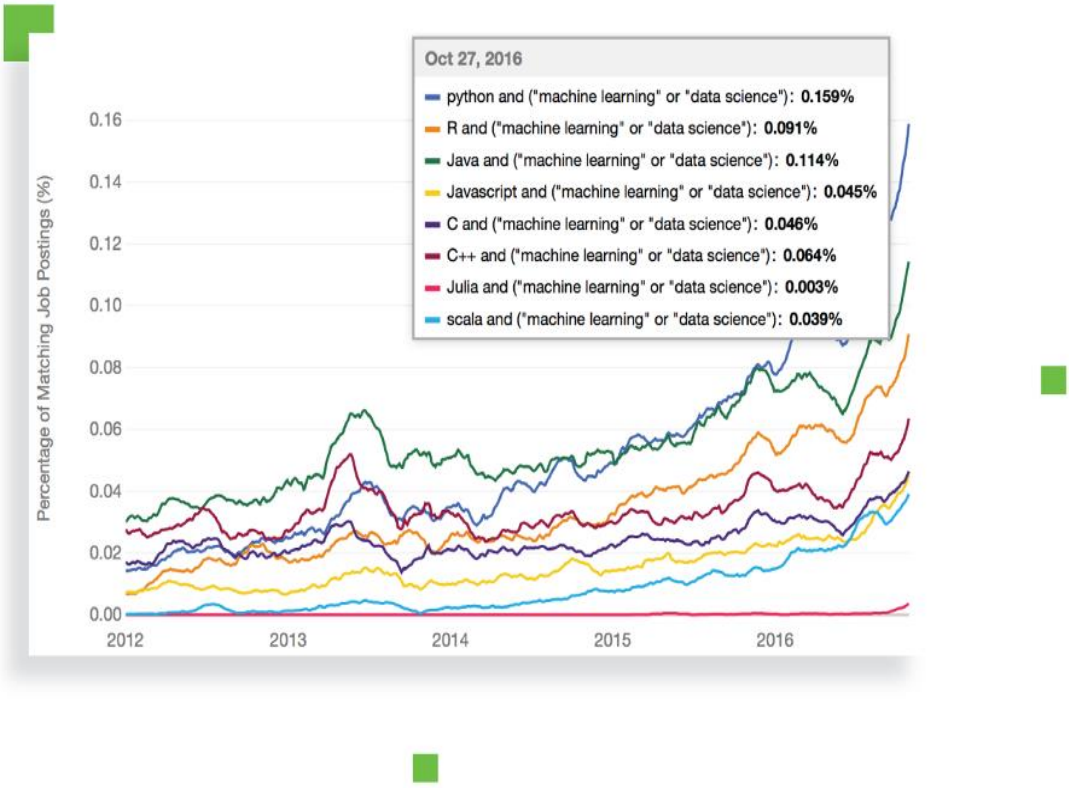

Figure 2.4: Python is the major code language for AI and ML. [21]

2.4.2 Why Python?

- **Many useful libraries**

In addition to the simple code that can type (Syntax), Python has a large collection of libraries that can be embedded and used to make application more efficient and capable of performing a host of additional tasks [23]. Such:

- Scikit-learn library: it one of machine learning libraries in Python contains many algorithms and methods used in machine learning such as Classification, Clustering and Regression, as well as for data processing and model evaluation, Scipy, Numpy, Matplotlib and many other libraries. The scikit-learn library focuses on data modeling and does not focus on data loading, handling and summarizing, which is the primary role of the Pandas and Numpy libraries [24].

- Panda library: is an open source library that provides high performance and ease in structure and analysis of data in its use as a library with Python software. The library serves the financial, statistical, social and engineering fields. The library works well with incomplete, unorganized and unnamed data and an example (any type of data that is likely to exist in the real world). The library provides tools for forming, merging, reshaping and dissecting a set of data [25].

- Numpy library: using as a foundation library for scientific computing with Python and contains other things so that the basic input and output matrices are used. In short, the object introduces different dimensions of matrices as well as procedures that enable developers to perform mathematical and statistical functions with as little code as possible, Linearity and finding random numbers and other advanced functions in the mathematical and statistical fields [26].

- Matplotlib library: The Matplotlib library using in Python for two-dimensional planning and graphs. The library is beautiful but low-level, meaning it needs a lot of code to create beautiful graphics. However, the library is flexible enough with commands to create any graphic format you want [27].

- NTLK library: The library is a leader in building applications that are used to understand human language data in Python as they are designed with NLP algorithms. The basic functions of the library allow you to understand and distinguish the text, identify the semantic data and display the analysis trees, which are similar to the outline of sentences that reveal parts of speech and dependencies. From here you can do more complex things like emotion analysis and automatic summary [28].

- **No barrier**

This feature in python helps many developers to quickly start learning python and start their application without barriers, and without spending a lot of time in learning the syntax; beside the plenty of documentation which affordable to learn the language, and the online wide community which support any project.

- **Flexibility**

It offers an alternative to pick either to utilize OOPs or scripting. There's likewise no compelling reason to recompile the source code, engineers can actualize any progressions and rapidly observe the outcomes. Software engineers can join Python and different dialects to achieve their objectives.

- **Platform Independence**

The python libraries for machine learning can run on many operating systems such: MacOS, Linux, Windows., Etc.

The developers using PyInstaller to install python on different platforms, which in turn reduce cost and time which using in test the application on platforms.

- **Readability**

Python is anything but difficult to peruse so every Python engineer can comprehend the code of their friends and change, duplicate or offer it. There are likewise devices like IPython accessible, which is an intuitive shell that gives additional highlights like testing, investigating, tab-finish, and others, and encourages the work procedure

3. METHODOLOGY

The machine learning problems follows the following model in the supervised learning, this model will be used in this research, where the tweets will be the training text documents, which from the feature vectors will extracted to be feed to the machine learning algorithms, in order to build predictive model able to predict the coming new tweet to which author follow: figure (3.1) shows [6]

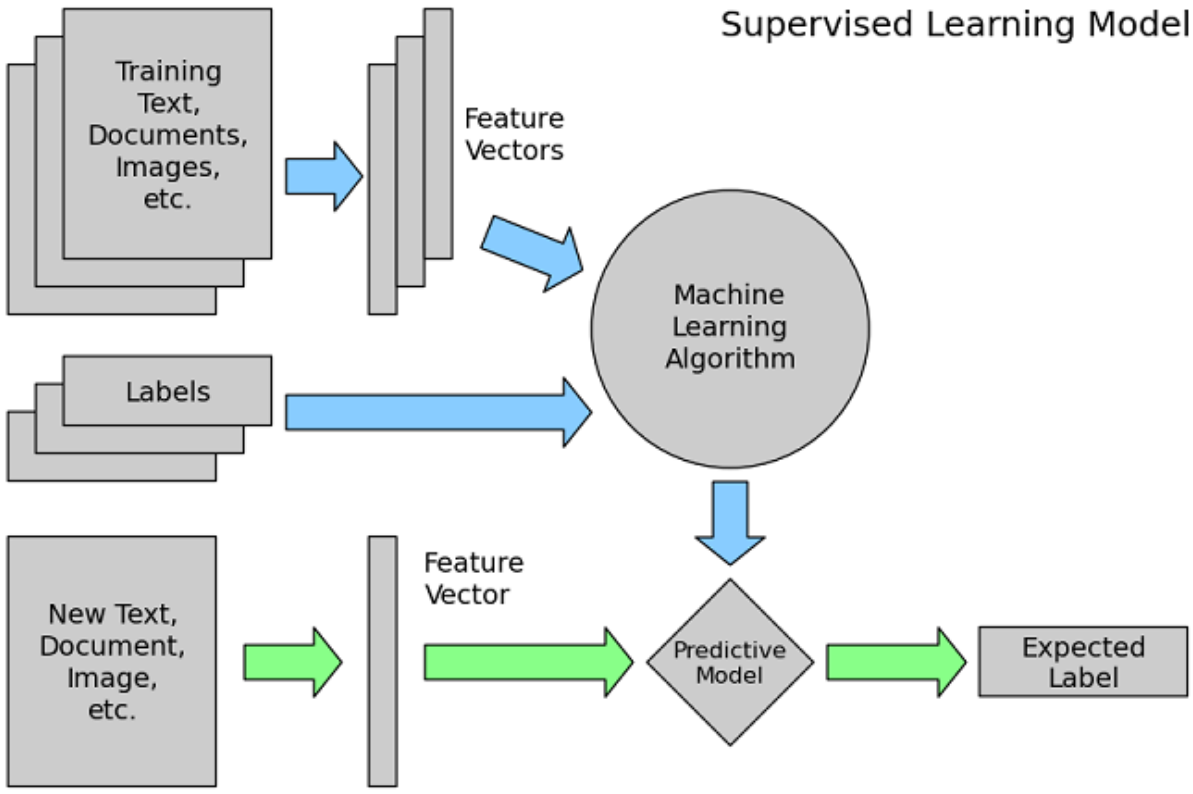

Figure 3.1: Supervised learning model [6]

The scope of the research will be as following:

3.1 COLLECTING A DATA SET FOR MANY AUTHORS

Nowadays Twitter considering is one of the most important social networks, which spread fast in the recent years. Twitter is a mini-service that allows two recipients to send text messages that do not exceed 140 characters per message, and may call for a condensed text for many details. This mini-service was introduced in early 2006 when the US Company Obbary undertook a development

research to serve micro-blogging. The company then made use of this service to the general public in October of the same year, and then took this site to spread as a modern service in the field of micro-blogging. After that, the company separated the mini service from the parent company and developed a new name for Twitter in April 2007.

Twitter is most popular social networking sites on as Facebook and is considered the biggest competitor to it, where it offers a mini service for users who are able to send e-mail as a message. 140 characters for a one-to-one message. Visitors can read and write replies in a timely fashion, the message is called tweet.

This section discuss the used method to collect the data from Twitter, where the tweets will be collecting using the REST API, which can be embedded in crawler, built in Python programming language, by it we can communicate and send the requests to Twitter via Twitter API, which is how we provide Twitter to access the data stored on the site [30].

3.1.1 Twitter API Structure Consists of Four Main Purposes

User: If he has Twitter profiles, he can follow others, create lists, tweet, send messages, reply to the others tweets and other services available on the site.

Tweet: it is the building key in Twitter site and also known as updating the status. It takes several instances: such being a post or a response to another user, with the possibility of deleting and adding to the user's preference.

Entities: Provide information about tweet content that is published on Twitter such as links, images, number of followers...

Places: it’s defined by name and coordinates that can be placed in the tweet.

3.1.2 Access to Twitter API

To access Twitter API, a developer account on twitter should be available, it can be created using the following link: https://dev.twitter.com ; The process of creating the application is characterized by simplicity, after creation the application and the registration process. The following four keys are provided by Twitter to access its API, as in the following table (3.1):

Table 3.1: Keys are provided by Twitter to access its API

Consumer Key (API Key)	iqln8PQoyZW5ne3uks2FGZ
Consumer Secret (API Secret)	INxJPTfcidiCmISvLL2sMJK83xoYe62Nd96pBstuwTJFMDJ
Access Token	207044069-NijfONAVIZsSv5M8DryeaaDW2pc7li9tg5io
Access Token Secret	3T8edZWzotmUoSneR4um2hhEhyfqkgsv7RgsUJ7hkk

These four objects are used to enable the connection between the application and API, where four variables are defined to secure the credentials and authentication process as following:

Twitter API credentials

- consumer_key = " iqln8PQoyZW5ne3uks2FGZ "
- consumer_secret = " INxJPTfcidiCmISvLL2sMJK83xoYe62Nd96pBstuwTJFMDJ"
- access_key = "207044069-NijfONAVIZsSv5M8DryeaaDW2pc7li9tg5io"
- access_secret = "3T8edZWzotmUoSneR4um2hhEhyfqkgsv7RgsUJ7hkk"

Then the connection is done using Access Token and Token Secret via TwitterAPI:

- Auth=tweepy.OAuthHandler("iqln8PQoyZW5ne3uks2FGZjHq","INxJPTfcidiCmISvLL2sMJK83xoYe62Nd96pBstuwTJFMDJ7YV ")
- auth.set_access_token("207044069-NijfONAVIZsSv5M8DryeaaDW2pc7li9tg5ionpjP","3T8edZWzotmUoSneR4um2hhEhyfqkgsv7RgsUJ7hkkQLf ")

- api = tweepy.API(auth)

The following figure (3.2) shows use case and how to handle a twitter site:

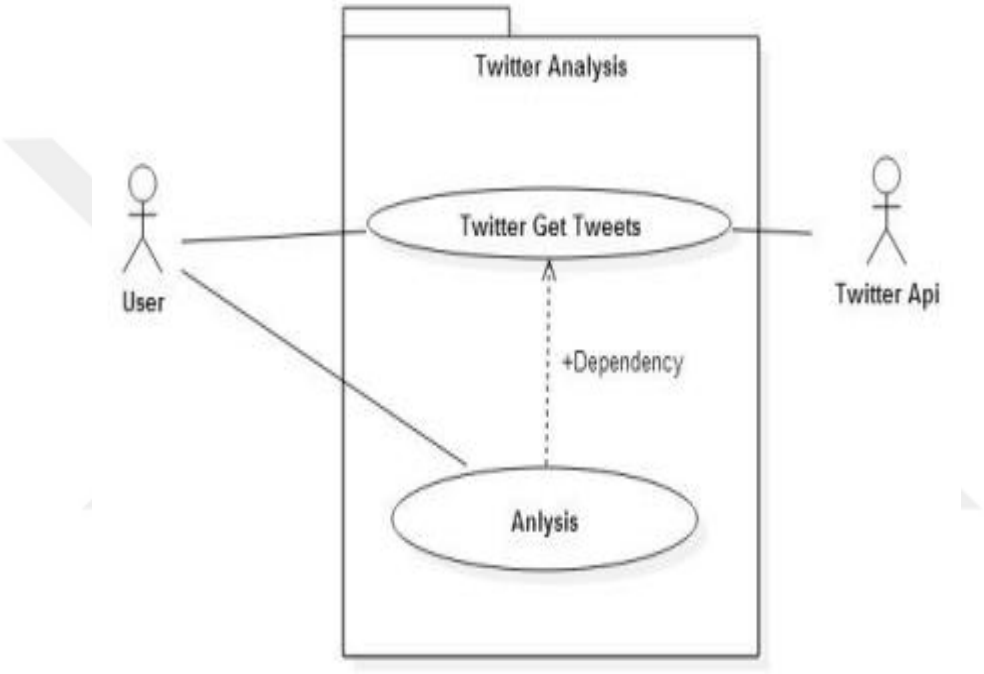

Figure 3.2: User connection with the Twitter API use case

The following table (3.2) shows the scenario to get tweets from twitter:

Table 3.2: Scenario to get tweets from twitter

get_all_tweets			
The main scenario	The developer enter the user_id in the function, for the user whose data we want to pull		
Actor	Twitter Api – User		
Pre-conditions	Available Internet connection + available access_tokens		
Events series	User	System	Twitter_API
	• Developer enter the user_id • The developer run the application to get the tweets	• The system opens a channel with the Twitter_API • The system send the Access_token which special for the application	• Twitter_API validate the access_token which provided by the sender • Return the requested data
Alternative scenario	• The access_token is invalid or expired • There is no internet connection • There is not data to be returned		

3.2 PREPROCESS THE DATA SET

It has been done in several stages: The process of cleaning the text including removing words that are outside the English language in addition to the special symbols, the removal of English stop words, the formation of N-gram length 2, And the numbers; regular expressions have been used in the cleaning process; Python programming language have been used in this step.

Regular Expressions :the literal meaning is typical expressions, based on the so-called Pattern, and can be referenced by regex or regexp. It is a method used to describe texts for searching and

matching within the text or sentences to compare them and then extract them or verify their validity, for example if you want to delete all the spaces in a file can work by manually searching the file and delete all the distances you find, You can build a script that contains a huge number of If and Loop conditions and make sure that each character is a distance or not. The last solution is using Regex. In one word, you can not exceed one line to do all of the above. And a tremendous effort especially in the Factor with large text, you can use this technique with many languages such as Java, PHP, Python, C, JavaScript and others.

The Regex technique consists of several symbols that you can use to implement your expression. To use it You can combine the preceding symbols to form a single typical expression such as [a-zA-Z] \ s [0-9], or you can separate them by groups using brackets [a-zA-Z] (\ s) 9]. Each group of previous groups can also be called by the $ symbol. For example, if you only want to use the first and third groups, call them by $ 1 $ 3 and so on.

there are alot of symbols can by used in regular expression, for example :

- The tag "^" means the first line
- The "$" sign means the last line
- The tag "|" Means "or" ... such or such
- the point "." Means any character (a letter, a number, a tag, and a corresponding "*" in Linux Shell)
- The tag "&" means the value of the search result. (For example if we search for "Mohammed" and replace it with "& Yen" the name will eventually appear "Mohammad")
- The tag "\" means if at the first word, for example, the expression "\ <Zaid" will match "Zidane" and will not find "Abozeid"
- The tag "\>" is like its predecessor but at the end of the word, the expression "Z> will match" Abuzid "and not" Zidane "
- Mark "?" Means there is at least one or none. For example, if we write "Ahd", it will match the word "Ahmed" and will also match the word "one."

- "*" Means no or infinite number of times
- The "+" sign means that it exists at least once or more times
- The mark "[^ ANY]" means the exclusion of the said after the sign "^" meaning that if we say "[^ Ahmed]" all words except Ahmed will be matched
- The tag "\ t" means tab
- The tag "\ s" means a space
- The tag "\ S" means any character other than space
- The tag "\ n" means a new line, that is, if you press Enter
- The tag "\ d" means any number
- The tag "\ D" means anything other than numbers
- "\ W" means any word (consisting of letters, numbers, or "_")
- The tag "\ W" means anything other than words (any other symbols)

İn this research the following preprocessing have been done:

- The Retweets have been deleted, because they dont reporesetn the author style in writing, whereas the research goal is to detect the authors from their writing style so that dont help the reseaerch goal
- The month names are converting using the regular expression to '\m'
- The websites adresses are converted to space, where exist of these adresses insdie the tweet dont express the writing style
- Converting the hashtags to word hash, the hashtags is a s a word preceded by #, which is used to classify your publications on Twitter, if the user want to publish an update related to the city of Tunisia, for example, you write in the update #Tunisia, the other use is to follow all the publications on Twitter about a particular classification, for example if you want to

follow updates about an event or issue you write in the search box #; For webmasters, Hashtag is an effective way to promote your page and reach more Facebook users. It is also a simple way to communicate and communicate with people who have common interests

- The symbol “%” are converted to \p
- The puctuation marks which used inside the tweet, have been after preprocessing !
- The imojes which used inside the tweet, have been after preprocessing ?

3.3 EXTRACTING THE FEATURES

There are many features could be extracted from the text, in the following the features which have been used in this research:

- Char N-Gram:

  N-grams are simply all combinations of adjacent words or letters of length n that you can find in your source text. For example, given the word fox, all 2-grams (or “bigrams”) are fo and ox. You may also count the word boundary – that would expand the list of 2-grams to #f, fo, ox, and x#, where # denotes a word boundary. The basic point of n-grams is that they capture the language structure from the statistical point of view, like what letter or word is likely to follow the given one. The longer the n-gram (the higher the *n*), the more context you have to work with. Optimum length really depends on the application – if your n-grams are too short, you may fail to capture important differences. On the other hand, if they are too long, you may fail to capture the “general knowledge” and only stick to particular cases.

As seen in the following figure (3.3) shows [31] the char N-Gram is about extracting N character sequence from the sentences, whatever N.

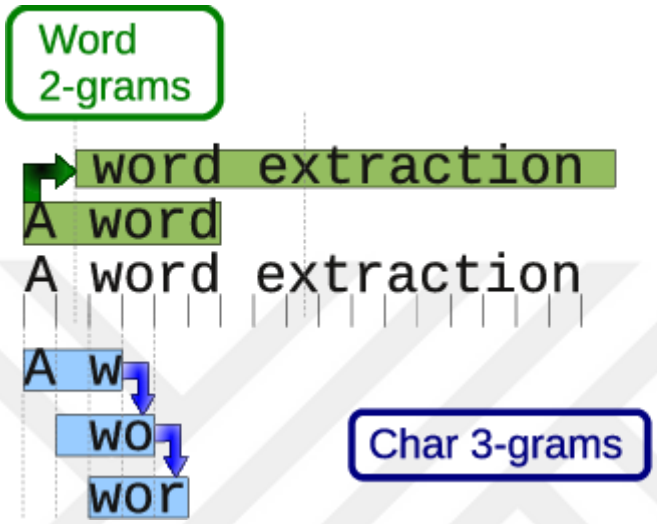

Figure 3.3: Char N-Gram [31]

- Word N-Gram:

  The word is about extracting the word N-Gram sequence from the text, whatever N As seen in the following figure (3.4) shows [32]

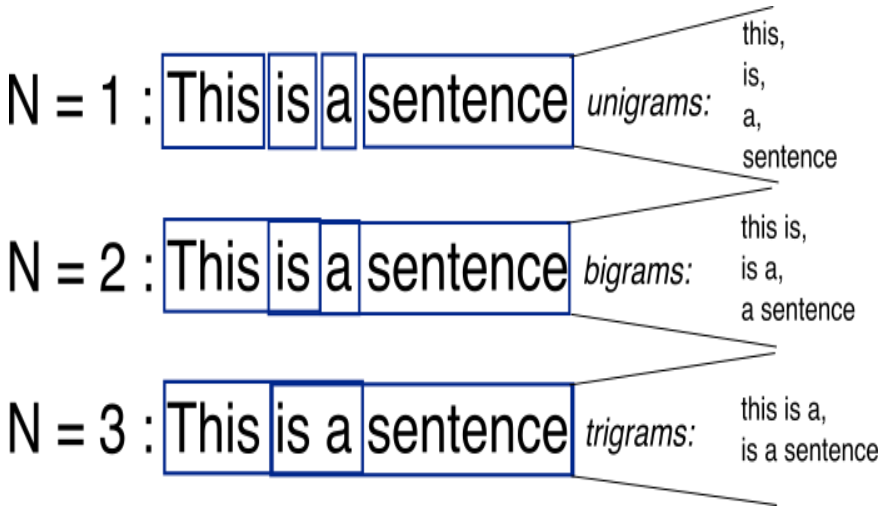

Figure 3.4: Word N-Gram [32]

- Stop words

  In the following figure (3.5) shows [33] list about all the stop words existed in the English language.

```
> stopwords("english")
  [1] "i"          "me"         "my"         "myself"     "we"
  [6] "our"        "ours"       "ourselves"  "you"        "your"
 [11] "yours"      "yourself"   "yourselves" "he"         "him"
 [16] "his"        "himself"    "she"        "her"        "hers"
 [21] "herself"    "it"         "its"        "itself"     "they"
 [26] "them"       "their"      "theirs"     "themselves" "what"
 [31] "which"      "who"        "whom"       "this"       "that"
 [36] "these"      "those"      "am"         "is"         "are"
 [41] "was"        "were"       "be"         "been"       "being"
 [46] "have"       "has"        "had"        "having"     "do"
```

Figure 3.5: Stop words in English [33]

- Bag Of Words BOW:

  The bag of words is a victories representation for the data, as shown in the following figure (3.6) shows [34].

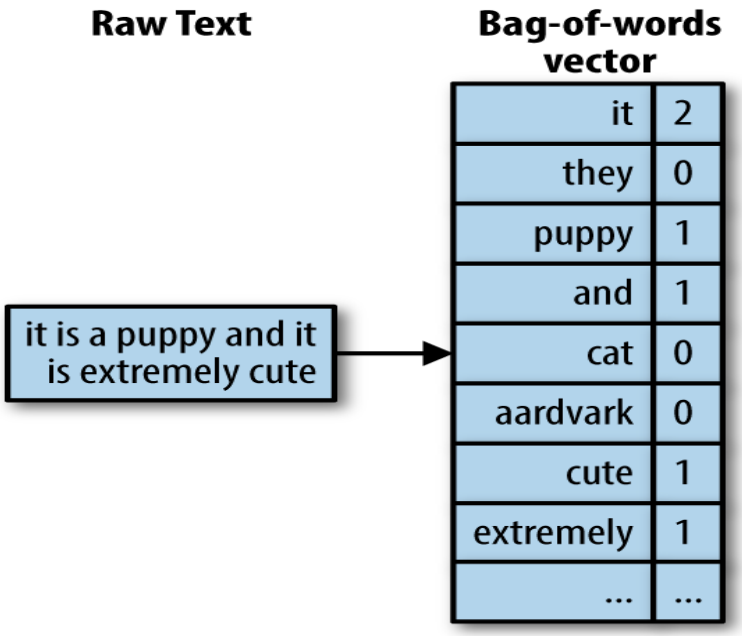

Figure 3.6: Bag of Words representation [34]

- TF-IDF

  According to Bapi “TF-IDF stands for term frequency-inverse document frequency, and the TF-IDF weight is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Variations of the TF-IDF weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query .One of the simplest ranking functions is computed by summing the TF-IDF for each query term; many more sophisticated ranking functions are variants of this simple model” [35]

- Lexical features

  It’s about manipulating the tweets from the lexical viewpoint, which care about the words in the English language and the relation between them

- Punctuation features:

  There are many punctuation marks in the embellish language such [36]:

  - the period (or full stop in British English)
  - the comma
  - the exclamation mark
  - the question mark
  - the colon
  - the semicolon
  - the quotation mark
  - the apostrophe
  - the hyphen and the dash
  - parentheses and brackets

In this research just the exclamation mark, the question mark and the apostrophe, Will be used as a punctuation features.

3.4 BUILD THE CLASSIFICATION MODEL

As mentioned in the previous chapter, there are many classification' types, to make the classification to the data, a classification algorithm should be use, there are a lot of classification algorithms has been developed, this research use three supervised learning algorithms to classify the data, and compare the result between them:

3.4.1 SVM

SVM Is a technology of machine learning techniques under the supervision used for data classification. Classification has wide applications in biology, for its high accuracy. Nor does it need to have a deep understanding of mathematical theoretical matters behind them. It is also easier than neural networks. Neural Networks the classification process in SVM includes two phases. The first is the training phase, the second the testing phase. In the training phase, a set of data called Training Dataset is provided. This group includes Instances. Each case contains two parts: the first is the Class. The second is the Attributes. The objective of the SVM is to produce a model based on data Training so that Class can be given for descriptive cases only [37].

The main concept in the SVM is:

Hyperplane:

Supposing have a set of training data. Each instance of Instance represents a text file that contains an article. This can be about Java or design models. The first represents the repetition of the word "row," and the second represents the repetition of the word "method." Each case can be either "Java" or "Design Patterns." These situations represent a ray in binary space, each axis representing the shape. Green dots are rays that represent Java and red dots rays represent design model articles. You can notice that Java articles meet in the upper-left and Design Model articles meet in the lower-right corner. as the following figure (3.7) shows [10]: These rays can be separated by line separating them. In this case, it is possible to predict the class of a new article that is unknown to the class by identifying its location from this space (on any side of the dividing line). Make the label of this

super-surface line. This is called a Hyperplane because it is a surface in the larger dimensions of the two as the following figure (3.8) shows [10]:

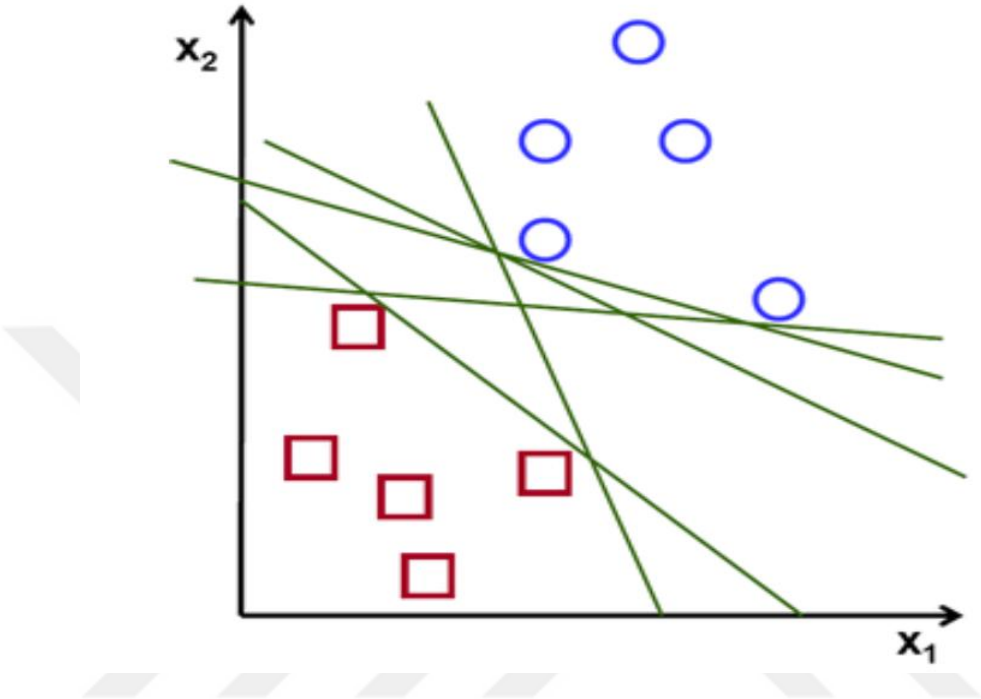

Figure 3.7: Representing the java and pattern design articles [10]

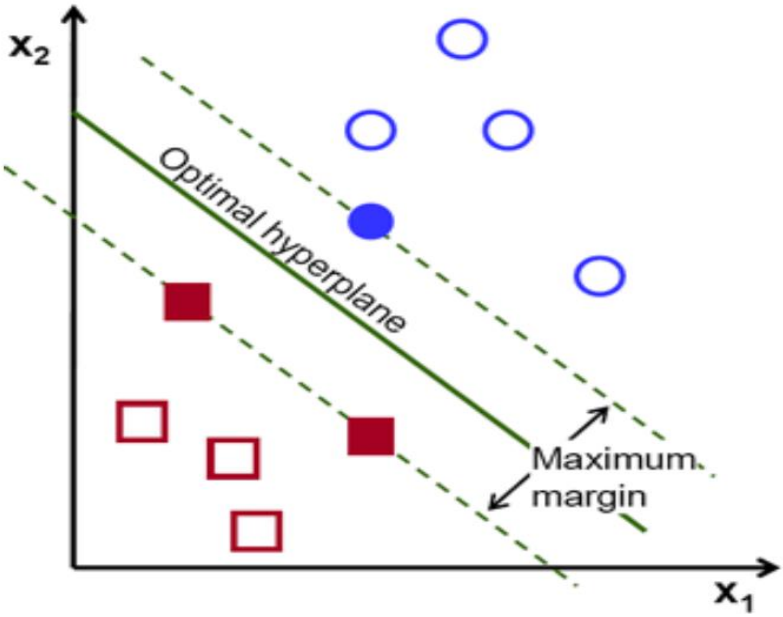

Figure 3.8: The Hyperplane concept in SVM [10]

The SVM algorithm also using in multi classification problem, where the goal is creating the Hyperplane to separate between many classes, as the following figure (3.9) shows [38]:

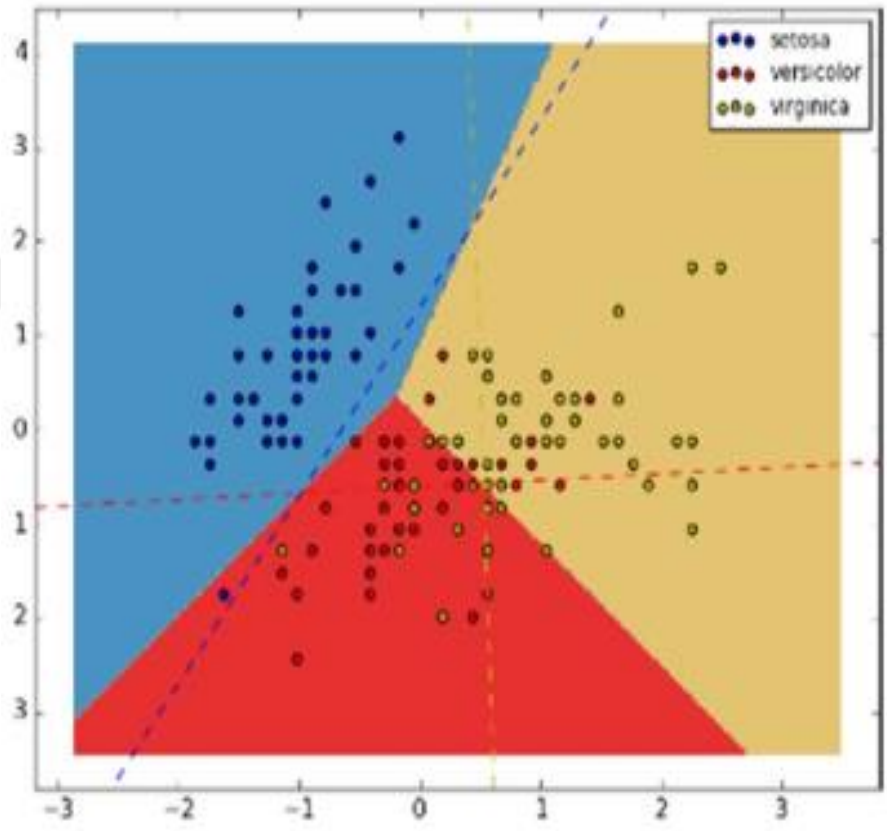

Figure 3.9: Multi-class SVM on three classes [38]

3.4.2 Logistic Regression

"Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).

In logistic regression, the dependent variable is binary or dichotomous, i.e. it only contains data coded as 1 (TRUE, success, pregnant, etc.) or 0 (FALSE, failure, non-pregnant, etc.).

The goal of logistic regression is to find the best fitting (yet biologically reasonable) model to describe the relationship between the dichotomous characteristic of interest (dependent variable =

response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a logit transformation of the probability of presence of the characteristic of interest" in eq (3.1) as following Figure (3.10) shows [39]:

$$\mathbf{logit(p) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_kx_k} \tag{3.1}$$

Where p is the probability of presence of the characteristic of interest. The logit transformation is defin"ed as the logged odds in eq (3.2):

$$odds = \frac{p}{1-p} = \frac{probability\ of\ presence\ of\ characteristic}{probability\ of\ absence\ of\ characteristic} \tag{3.2}$$

And in eq (3.3):

$$logit(p) = ln\frac{P}{1-p} \tag{3.3}$$

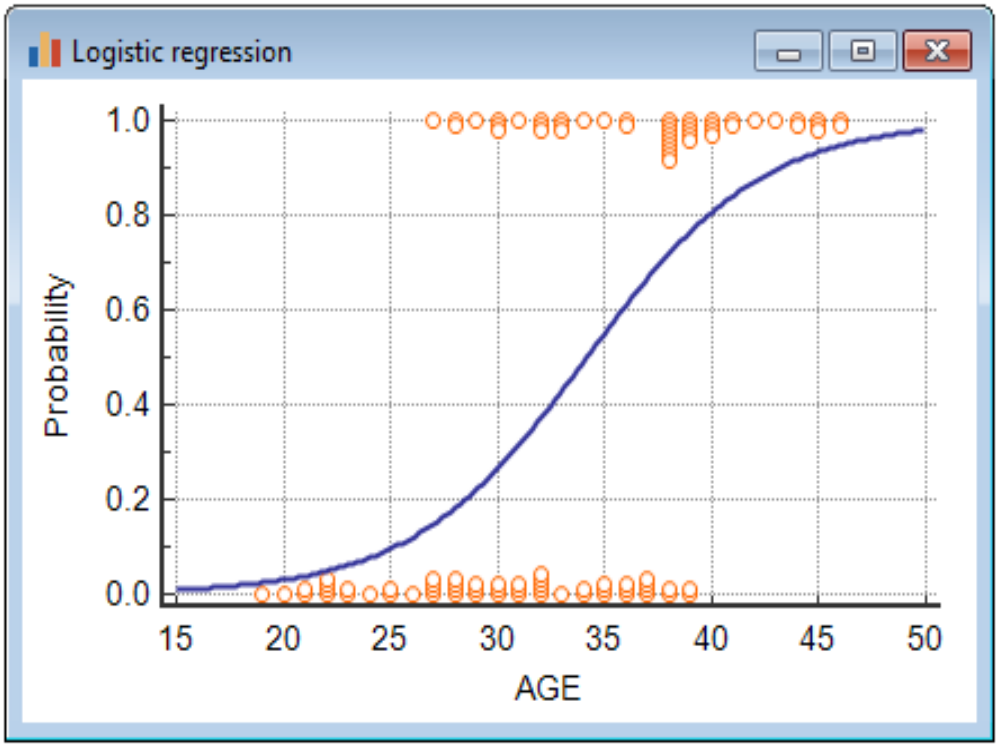

Figure 3.10: Logistic regression representation [39]

3.4.3 Random Forest

According to Leo Breiman: "Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges a.s. to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. Using a random selection of features to split each node yields error rates that compare favorably to Adaboost but are more robust with respect to noise. Internal estimates monitor error, strength, and correlation and these are used to show the response to increasing the number of features used in the splitting. Internal estimates are also used to measure variable importance. These ideas are also applicable to regression" as following Figure (3.11) shows [41]:

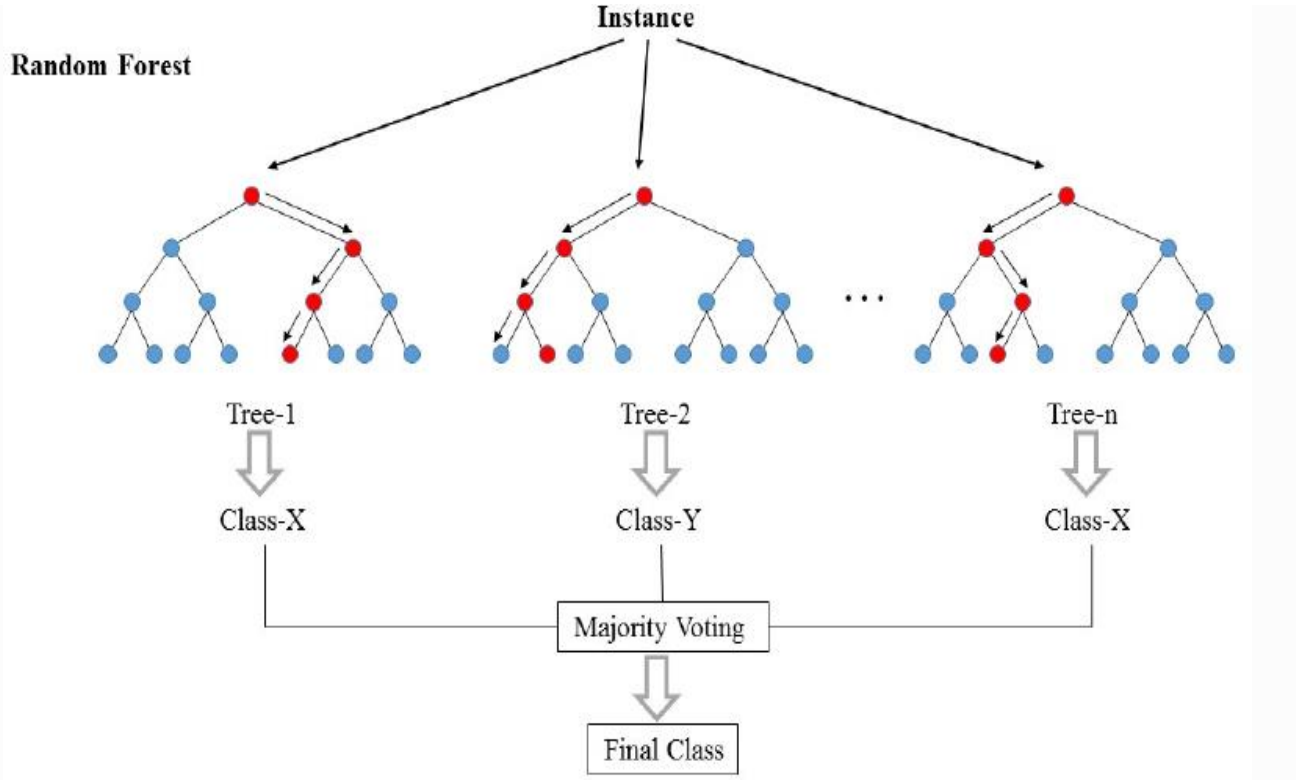

Figure 3.11: Random forest algorithm hierarchy [41]

3.5 EVALUATION MEASURES

Before implementing the previous three classification algorithms on the dataset, an evaluation measures should be specified to verify the result, in the following some evaluation measures, which frequently using in the literatures to evaluate the classification algorithms:

At first it's good to define the confusion matrix, which is the base in the evaluation measures, the confusion matrix determine the number of the classes that the algorithm succeed or failed in predict them, as the following figure (3.12) shows [42]:

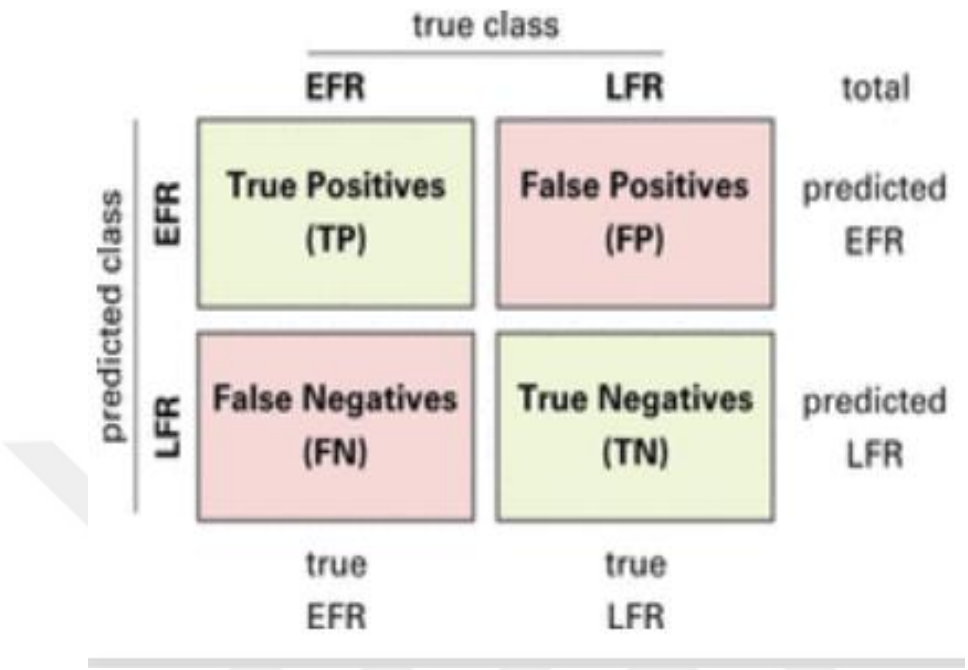

Figure 3.12: A Confusion matrix from a two-class classification problem [42]

The confusion matrix contains the following items (True Positive, False Positive, True Negative, and False Negative), which defined by Kaur as following [12]:

True Positive (TP): “is the correct recognition of a tweet written by the user i.e. a tweet actually written by the user is detected to be by the same user"

False Positive (FP): “is the false recognition of the tweets written by an outsider to be by the user in question”

True Negative (TN): “defines the case of correct recognition of a tweet written by some other user i.e. a tweet written by an outsider is recognized to be by an outsider”

False Negative (FN): “is falsely recognizing a tweet written by the genuine user to be by an imposter”

And the following measure which depend on the above items will be calculated to evaluate the classification algorithms:

Recall

This measure is using to determine the percentage of the true positive expected tweets on the total of the true positive adding to the false negative expected tweets [12], and calculated using the following formula in eq (3.4):

$$Recall\ (R) = \frac{TP}{TP + FN} \tag{3.4}$$

Precision

This measure is using to determine the percentage of the true positive expected tweets on the total of the true positive adding to the false positive expected tweets [12], and calculated using the following formula in eq (3.5):

$$precision\ (p) = \frac{TP}{TP + FP} \tag{3.5}$$

F1_score

This measure is determining by the previous two measures [12], and calculated using the following formula in eq (3.6):

$$F_{score} = \frac{2 * precision * Recall}{Precision + Recall} \tag{3.6}$$

3.6 EXTRACT THE RESULT

In the following the implementation accuracy measures for each supervised classification algorithm according to the extracted dataset:

Logistic Regression Algorithm

In this section, the results for the implementation of the logistic regression algorithms have been got, as following Table (3.3):

Table 3.3: Logistic regression implmentation measures

Number of extracted features	108980
Model Runtime	0.58 Minutes
F1 score	[0.84778013, 0.80054645, 0.79558011, 0.7114094, 0.78397711]
Recall	[0.90315315, 0.80494505, 0.79120879, 0.6437247, 0.77183099]

Classification Report as following Table (3.4):

Table 3.4: Logistic Regression classification report

	Precision	**recall**	**f1-score**	**support**
0	0.80	0.90	0.85	444
1	0.80	0.80	0.80	364
2	0.80	0.79	0.80	364
3	0.80	0.64	0.71	247
4	0.80	0.77	0.78	355
Average / total	0.80	0.80	0.80	1774

Confusion Matrix as following Table (3.5):

Table 3.5: Logistic Regression confusion matrix

401	14	14	9	6
26	293	13	17	15
21	18	288	5	32
37	26	8	159	17
17	17	37	10	274

SVM algorithm:

In this section, the results for the implementation of the Support Vector Machine algorithm have been got, as following Table (3.6):

Table 3.6: SVM implementation measures

Number of extracted features	108980
Model Runtime	0.38 Minutes
F1 score	[0.83014862 0.75690608 0.76944444 0.63793103 0.75358166]
Recall	[0.88063063 0.75274725 0.76098901 0.59919028 0.74084507]

Classification Report as following Table 3.7):

Table 3.7: SVM classification report

	Precision	**recall**	**f1-score**	**support**
0	0.79	0.88	0.83	444
1	0.76	0.75	0.76	364
2	0.78	0.76	0.77	364
3	0.68	0.60	0.64	247
4	0.77	0.74	0.75	355
Average / total	0.76	0.76	0.76	1774

Confusion Matrix as following Table (3.8):

Table 3.8: SVM confusion matrix

391	17	13	14	9
31	274	15	29	15
20	17	277	15	35
39	30	9	148	21
17	22	42	11	263

Random Forest Algorithm

In this section, the results for the implementation of the Random Forest algorithm have been got, as following Table (3.9):

Table 3.9: Random Forest implementation measures

Number of extracted features	108980
Model Runtime	7.71 Minutes
F1 score	[0.73272727 0.69333333 0.66473988 0.48837209 0.65128901]
Recall	[0.90765766 0.64285714 0.63186813 0.34008097 0.67605634]

Classification Report as following Table (3.10):

Table 3.10: Random Forest classification report

	Precision	**recall**	**f1-score**	**support**
0	0.61	0.91	0.73	444
1	0.75	0.64	0.69	364
2	0.70	0.63	0.66	364
3	0.87	0.34	0.49	247
4	0.63	0.68	0.65	355
Average / total	0.70	0.67	0.66	1774

Confusion Matrix as following Table 3.11):

Table 3.11: Random Forest confusion matrix

403	15	7	4	15
71	234	24	5	30
65	16	230	3	50
69	31	16	84	47
48	15	51	1	240

4. CONCLUSION

This research discussed the ability to predict the author of specific message between many authors, the feasibility of this research discussed previously; due of that a classification algorithm able to make multi classification (which mean assigning a tweet to one from many authors) used, many supervised learning algorithms have been used, the algorithm which achieved the highest prediction score comparing with other used machine learning algorithms is the “Logistic Learning” algorithm which as mentioned in its definition its measure the occurrence probability for specific instance; the algorithm here gave average to 80%, which mean that the algorithm can by this percentage say this tweet is written by its author .

As a future work more features can be combined with these features to rise the prediction accuracy, or a new classification algorithm can be tested.

REFERENCES

[1] Stamatatos, E. (2009), “A survey of Modern authorship attribution methods”, Journal of the American Society for Information Science and Technology, 538-556.

[2] Metin TEKKALMAZ. Yiğithan DEDEOĞLU.(2006),»Authorship Attribution», CS533-znformation Retrieval Systems, www.cs.bilkent.edu.tr/~canf/CS533/.../Authorship%20 Attribution.ppt

[3] Shaker, Kareem. (2012). »Investigating Features and Techniques for Arabic Authorship Attribution», PhD. Thesis, Heriot-Watt University, Department Of Computer Science School of Mathematics and Computer Science, March 2012.

[4] Stamatatos, E. (2008). «Author identification: Using text sampling to handle the class balance problem», Information Processing and Management, 44(2), 790-799.

[5] Rocha, Anderson, et al. "Authorship attribution for social media forensics." IEEE Transactions on Information Forensics and Security 12.1 (2017): 5-33.

[6] Brocardo, Marcelo Luiz, Issa Traore, Sherif Saad, and Isaac Woungang. "Authorship verification for short messages using stylometry." In Computer, Information and Telecommunication Systems (CITS), 2013 International Conference on, pp. 1-6. IEEE, 2013.7

[7] Castro, A., & Lindauer, B. (2012). Author Identification on Twitter.

[8] Rabab'ah, A., Al-Ayyoub, M., Jararweh, Y., & Aldwairi, M. (2016, November). Authorship attribution of Arabic tweets. In 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA) (pp. 1-6). IEEE.

[9] Rao, O. S., Raju, N. G., Srilalitha, Y., & Bharathi, M. P. Authorship Attribution using Unsupervised Clustering Algorithms on English C50 News.

[10] Rocha, A., Scheirer, W. J., Forstall, C. W., Cavalcante, T., Theophilo, A., Shen, B., & Stamatatos, E. (2017). Authorship attribution for social media forensics. IEEE Transactions on Information Forensics and Security, 12(1), 5-33.

[11] Usha, A., & Thampi, S. M. (2017, December). Authorship Analysis of Social Media Contents Using Tone and Personality Features. In International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage (pp. 212-228). Springer, Cham.

[12] Kaur, R., Singh, S., & Kumar, H. (2018). AuthCom: Authorship verification and compromised account detection in online social networks using AHP-TOPSIS embedded profiling based technique. Expert Systems with Applications, 113, 397-414.

[13] Robbins, I. P. (2012). Writings on the Wall: The Need for an Authorship-Centric Approach to the Authentication of Social-Networking Evidence. Minn. JL Sci. & Tech., 13, 1.

[14] http://scikit-learn.org/stable/modules/multiclass.html

[15] http://www.nilsschaetti.ch/2018/01/23/short-history-artificial-intelligence-nn/

[16] Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. Behavioral and brain sciences, 40.

[17] https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861

[18] Mohammed, M., Khan, M. B., & Bashier, E. B. M. (2016). Machine learning: algorithms and applications. Crc Press.

[19] https://pythonbasics.org/7-reasons-to-learn-python/

[20] https://sensip.engineering.asu.edu/research/image-and-video-processing/

[21] https://www.coursera.org/learn/machine-learning-with-python

[22] https://www.ibm.com/developerworks/community/blogs/jfp/entry/What_Language_Is_Best For_Machine_Learning_And_Data_Science?lang=en

[23] https://packaging.python.org/overview/

[24] https://scikit-learn.org/stable/

[25] https://pandas.pydata.org/

[26] https://www.numpy.org/

[27] https://matplotlib.org/

[28] https://www.nltk.org/

[29] https://www.catadoptionteam.org/

[30] https://dev.twitter.com/overview/documentation

[31] Béchet, N., & Csernel, M. (2012). Comparing Sanskrit Texts for Critical Editions: the sequences move problem. Polibits, (45), 27-35.

[32] http://cavajohn.blogspot.com/2013/05/how-to-sentiment-analysis-of-tweets.html

[33] Gadidov, B., & Priestley, J. L. (2018). Does Yelp matter? Analyzing (and guide to using) ratings for a quick serve restaurant chain. In Guide to Big Data Applications (pp. 503-522). Springer, Cham.

[34] http://uc-r.github.io/creating-text-features

[35] Bapi, R. S., Rao, K. S., & Prasad, M. V. (Eds.). (2019). First International Conference on Artificial Intelligence and Cognitive Computing: AICC 2018 (Vol. 815). Springer.

[36] https://www.ef.com/ca/english-resources/english-grammar/punctuation/

[37] Noble, W.S., What is a support vector machine? Computational Bioloogy, 2006. 24.

[38] Sarkar, D. (2016). Text Analytics with python. Apress.

[39] https://www.medcalc.org/manual/logistic_regression.php

[40] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[41] http://www.nrronline.org/article.asp?issn=1673-5374;year=2018;volume=13;issue=6;spage=962;epage= 970;aulast=Dimitriadis

[42] Bittrich, S., Kaden, M., Leberecht, C., Kaiser, F., Villmann, T., & Labudde, D. (2019). Application of an interpretable classification model on Early Folding Residues during protein folding. BioData mining, 12(1), 1.