Arab American University

Faculty of Graduate Studies

Semantic Analysis of Social Networks Using Hybrid System: Dictionary-Based and Fuzzy Logic

By

Ateeq Omer Abdulhadi Ateeq

Supervisor

Dr. Amjad Rattrout

This thesis was submitted in partial fulfillment of the requirements for the Master's degree in Computer Science.

February 2019

Semantic Analysis of Social Networks Using Hybrid System: Dictionary-Based and Fuzzy Logic

By

Ateeq Omer Abdulhadi Ateeq

This thesis was defended successfully on February 25, 2019 and approved by:

Committee members	Signature
1. Supervisor: Dr. Amjad Rattrout	.....
2. Internal Examiner: Pof.Dr. Mohammed Awad	.....
3. External Examiner: Dr. Mohammad Al-Dasht	.....

Declaration

I hereby declare that this master thesis has been written only by myself without any assistance of any third party and describes my own work unless otherwise acknowledged in the text of the thesis.

All references, verbatim extracts and information source are quoted and cited properly. Thus, I confirm that no source has been used in this thesis other than those indicated in the thesis itself.

This master thesis has not been accepted in any other previous application, in whole or in part for any degree.

Signature:
Ateeq Omer Abdulhadi Ateeq

Acknowledgment

I thank God the Almighty for sustaining me until the completion of this work.

I would like to express my love and gratitude to my beloved family for their understanding & endless love, through the duration of my studies.

I would also like to express my sincere gratitude to my supervisor Dr. Amjad Rattrout who was abundantly helpful and offered invaluable assistance, support and guidance through my Master study and research.

Abstract

In recent years, social networks become an information goldmine provides analyzes and inferences rich environment which can be exploited for the development of knowledge in various fields. The semantic analysis of social media can be classified into three main approaches which are content-based semantic analysis, user-based semantic analysis, and network-based semantic analysis. The first approach is concerned with the content of the posts and mainly on the textual context. The second approach is concerned with the social network users, using the user-based analysis is employing the categorizing of the users according to their patterns usage and the personal trend to have a user-based semantic analysis. The third approach is concerned with user network data such as friends, followers, followees, likes, and shares. In this research, our focus is on the first approach.

Several algorithms were used to reach the maximum possible accuracy in the semantic analysis of social networks; the most accurate results were obtained by using the dictionary based and the fuzzy logic algorithms. In this thesis, we worked to obtain better results by creating a hybrid system that fuses the dictionary based and the fuzzy logic to obtain better results rather than using each one of them independently.

As a conclusion of the results, we end with a prototype that calculates the polarities of the collected sentences and classify them into seven categories which are Very Positive, Positive, Good, Neutral, Not Good, Negative, and Very Negative in continuous learning manner, the prototype is learning from the previously collected data and changes its previous classifications, which was proven in the results mathematically.

Table of content

List of tables

List of figures

List of Abbreviations

Abbreviations	Description
AI	Artificial Intelligence
ASP	Active Server Pages
KDD	knowledge discovery from databases
SVM	Support Vector Machines
NLP	Natural Language Processing
POS	Part of Speech
CAAS	Classical Arabic Analysis System
URL	Uniform Resource Locator
NL	Natural Language
HTML	Hyper Text Markup Language
MVC	Model View Controller
API	Application Programming Interface
UTF	Unicode Transformation Format
JSON	JavaScript Object Notation
GMP	Generalized Modus Ponens
CPU	Central Processing Unit
GHz	Gigahertz
RAM	Random Access Memory
GB	Gigabytes
MS	Millisecond

Chapter 1: Introduction

1 Introduction

Realizing the emotions is a key feature for individual growth, development, and evolution. In addition to being signed for the improvement of Artificial intelligence, emotion analysis is also significant for polarity detection. The chance to robotically capture the public's opinions in fields such as social, political, marketing, and products has a growing interest in the scientific community, for its challenges, and the commercial world, for the amazing results in marketing and financial market forecasting. This has run to the developing areas of effective automation and sentiment analysis, which influence human-computer collaboration and information retrieval for refining people's opinions from the ever-growing online social data [1].

The effect mined from social media is identified to be reflective of communal behavior and opinion inclinations. The outcomes produced by the analysis of communal attitudes are gripping and specify that precise communal attitude indicators could be pulled out from online sources. Sentiment analysis has been used as an irreplaceable method for analyzing data for several objectives from social media sites. Via openly available online data to do sentiment analyses needs much less cost in terms of effort and time that would be required to manage large communal reviews, questionnaires, and surveys [2].

Sentiment words are used in several sentiment classification processes. These words are also known as “opinion words”. Sentiment words are permanently grouped into two arrangements based on their orientation: positive or negative. For example, “bright” has a positive sentiment and “unlucky” has a negative sentiment. There are three approaches to build a sentiment lexicon which are manually construction, corpus-based, and dictionary-based methods [3]. While the manual is hard, exorbitant, and time-consuming process, and corpus-based is more often used for sentiment analysis on documents, we choose the dictionary-based method as a base for the proposed system.

Unlike the traditional approaches that are depending on the words’ polarities only, we build a system that addresses the multi-domain sentiment analysis problem by applying fuzzy set theory like in [4].

1.1 Motivation

Sentiment analysis in itself is an important method to obtaining the public opinion in many fields; for the Arabic language, there is limited accountable researches that are aiming to provide a reliable sentiment analysis especially on social media as a data source.

We propose a sentiment analysis system that is mainly focusing on classifying pages' posts into several groups based on their content and the overall orientation in the page itself by applying the Fuzzy Logic model enhanced with the dictionary method for obtaining word-level polarities.

1.2 Problem Definition

Sentiment analysis is a complex process; complexity exists in many fields such as data collection, data processing, and data nature [5]. The complexity is increased when dealing with non-Latin languages due to the language structure and the lack of support tools [6].

We choose the Classical Arabic language as a target for this thesis; also, we choose Twitter and Facebook as data sources.

There are several challenges for processing Classical Arabic text which are the nature of the language and its richness in terms of words and phrases, the lack of existed data collection and cleanup tools, and the use of dialectal terms with Classical Arabic terms; in addition, there are some other challenges that are valuable to be mentioned such as the lack of Arabic polarities dictionary, the effect of some words on other words meaning, and the sarcasm effect.

Many references like [7] and [8] mentioned that sarcasm is a very complex problem even in Latin languages, we will not consider it as an objective in this project and we will focus on the other problems mentioned before.

1.3 Contributions

In this research, we propose a solution for the needs of sentimentally analyzing Classical Arabic text from social media. We showed the system components, models, and the interactions between them. The proposed model consists of different types of sub-models, each of them has a specific job and they communicate with each other in order to achieve the main system goal. We also proposed a new technique to classify the collected sentences into several categories in a continuous learning model.

In summary, the contributions of this thesis are:

- Classifying social media sentences into several groups which are Very Positive, Positive, Good, Neutral, Not Good, Negative, and Very Negative according to their polarities.
- Measuring the affection of the sentence popularity on the sentence polarity according to the same account's posts popularity.

- Measuring the affection of the sentences popularity on other sentences polarities for the same account.
- Measuring the affection of the account orientation on the sentence polarity for the same account.
- Measuring the change in posts classification during gathering posts and the effect of new posts on the previously collected posts polarities and classifications.

1.4 Thesis Objectives

The objectives of this research are summarized as follow:

1. Proposing an approach to classify Classical Arabic content in social media based on the polarity.
2. Measuring the effect of the social activities which are the likes and the overall account orientation, on the collected sentences classification.

1.5 Thesis Structure

This thesis consists of six chapters. Chapter one provides a brief introduction to the research and discussed the motivation, research problem, our contribution, and the research methodology. Chapter two describes the background of the research and topics that should be studied to get involved in the research details. Chapter three is the state of art literature review; it explains the related work in the following fields: dictionary approach, Fuzzy Logic approach, sentiment analysis on social media, and lexicon analysis. In chapter four, the methodology and the architecture of the proposed system were highlighted in details. In chapter five, we present the results of the classification process, highlight in the continuity of learning in the system and end the chapter with results discussion and explanations. Finally, we conclude in chapter six.

Chapter 2: Background

2 Background and Literature Review

2.1 Introduction

In this chapter, we are highlighting the background and literate review. We are covering the key related fields and the state-of-art articles that are needed to get involved with the research and proposed solution.

Text mining is a set of processes that aim to extract information from unstructured data [9], the extracted data could be in any field according to the targeted domain.

Sentiment analysis is considered a text mining method, it is done on different levels, these levels are word, sentence, and corpus; different techniques are used to enhance the sentiment analysis results such as SVM, Dictionaries, Fuzzy Logic and Classification techniques like Naïve Bayes [10]

2.2 Background

2.2.1 Sentiment analysis

Sentiment analysis (known also as opinion mining) is the use of language processing, text and biometrics to obtain and analyze information. It is heavily used in feedback analysis, decision making, and social media stuff.

The opinion mining fundamental task is to identify the polarity of a text if it is positive, negative, or neutral, furthermore, it can identify the feelings (emotions) in the text if it is angry, happy, sad ...etc.

Subjectivity and objectivity identification is a common classification task in opinion mining [10], while subjective depends on the context, the objective depends on the meaning of the text which is harder to detect [11].

The domain knowledge is one of the key issues in opinion mining to discover relations and obtain meaningful results; this is referred to as feature identification topic modelling [12].

2.2.2 Dictionaries -Based approach

The Dictionary-Based approach is used in sentiment analysis to obtain word level polarities. The approach is considered as a clear approach because is its quite direct approach and it is easy to implement and use.

The creation of a dictionary workflow starts with seeds, which are a small set of words that were collected manually and they have a pre-defined negative and positive polarities. The second phase is growing up the set by searching for seeds antonyms and synonyms, the latterly found

words are added to the dictionary. The next iterations start and the iterative operation ends when no more new words can be found. After the process completes, a manual check step is used to clean up the list [13].

2.2.2.1 SentiWordNet

SentiWordNet is a lexical resource for supporting sentiment classification and opinion mining applications; it is the result of the automatic annotation of all the WORDNET according to the notions of "positivity", "negativity", and "neutrality" [14].
SentiWordNet is giving the possibility to obtain the polarity for each word and classify the results in five different groups which are very positive, positive, negative, very negative and neutral.

2.2.3 Fuzzy Logic

Fuzzy logic is well-defined as a method of calculating to gain a degree of something rather than the old-style true or false "Boolean" logic. The idea was presented for the problem of machines understanding of natural language since it is hard to translate it to the terms of 0 and 1.

Fuzzy logic is used because it's obtaining rich linguistic results compared with the traditional sentiment analysis approaches. Therefore, sentiment analysis with the help of fuzzy logic will help the producers or consumers or any interested person for taking the efficient decision according to their products or services interest.

Fuzzy logic is used to obtain sentiments from a text or document. Fuzzy logic uses the concept of reasoning that gives outcomes in approximation rather than precise results and it is helpful in managing such approximate information. The numerical result of the sentence is evaluated between the ranges from 0 to 1. [15]

2.2.4 Text ontology

Ontology is defined as obvious, recognized requirements of common conceptualization of a domain of interest and it is restricted to that domain in terms of perceptions and relations [16].

Based on the ontology definition, the text ontology is the nature of the text for language, in terms of its basic categories and relations, as the ontology is often seen as the answer to the necessity for interoperable semantics in current information systems. Ontology learning from the text is the procedure of stemming high-level concepts and relations [17].

Text ontology is used in this thesis to apply to affection of a negation word or an affirmation word on another word; this ontological affection is discussed in section 3.3.2.4.

2.3 Literature Review

A complex question was the objective in [18] which was “How can morphological richness be treated in the context of Arabic Supervised Sentimental Analysis”; authors applied Sentiment and Subjective Analysis on Arabic language and they identify the dialects inside the text.

The main gap question which is “How to sentimentally analyze rich language like public Arabic” was the objective in [19]; authors applied supervised Sentimental Analysis on Arabic language and common Arabic language.

Both [18] and [19] provide a solid base to start from, but there are some issues that we still can improve like they ignore emotions, they do manual data collection, labelling, tagging and classification, also they were limited to the three basic categories (positive, negative and neutral).

In [20], an overview of studies on the sentiment classification was presented and two main topics were discussed, the sentiment classification approaches with respect to features, techniques, advantages and limitations, and the second topic was the used tools with respect to the different techniques used for sentiment analysis. Authors provided detailed analysis for the techniques and tools.

How to use Fuzzy logic for problems of Opinion Mining and Sentiment Analysis? This was the main gap that was covered in [21] the main reasons include flexibility, easy to understand and implement, relies on human-like expert knowledge and providing an excellent classification result.

[22] Proposed a survey that achieved a vision into many text mining methods applied to Arabic language, in addition to their implementation in the Holy Quran, sentiment analysis, and web documents. The latest improvement in the field of intelligent computing was explored by the survey.

2.3.1 Dictionary approach

In [10] authors presented the sentiment analysis approaches and methodologies related to Dictionary-based sentimental analysis, authors highlighted on sentimental analysis approaches, they discussed the Naïve Bayes, Machine learning, Dictionary-based, word level sentiment analysis and sentence level sentiment analysis, also they highlighted on sentiment analysis process phases and explained the sentiment classification performance metrics.

A new approach for sentimental analysis was presented in [23]; a comparative study of the sentiment analysis technique was also presented. The proposed approach classified reviews to positive, negative and normal based on a score that is calculated using SentiWordNet and WordNet

dictionary and by applying fuzzy logic. Authors presented a Hybrid system which gives high accuracy in the sentimental analysis by mixing the Dictionary-based with the Fuzzy Logic; the use of Fuzzy Logic is extremely helpful in sentiment analysis, but using only three categories is not enough especially for reviews.

Standalone based dictionary approach was used in [24] and [25]; in one hand, [24] highlighted on the usage of online dictionaries that is using the latest tweets that includes a specific word to calculate a word-level polarity extracted from sentence level polarity; on the other hand [25] used several dictionaries to obtain the final polarity by collaborating all the results together based on computational rules. There are two sides to judge the usage of online dictionary, from one side this is giving a real-time indicator for a word polarity, but on the other side, the polarity for the word is varying from a time to another which is an avoided issue when using an offline dictionary, also the online varying values are in a narrow range that should not highly affect the polarity results. The second research in [25] has two issues which are the manual division for the dictionary and the ignoring for the existed solutions and the usually used dictionaries like SentiWordNet.

In [26], [27], and [28], hybrid approaches that include the dictionary approach were proposed; the first and the third approaches in [26] and [28] respectively, proposed hybrid approaches which consisted of corpus and dictionary, and they were used to obtain the sentiment of tweets, while the second approach consisted of a dictionary and subjectively labelled tweets with labelled emoticons inside each tweet with using corpus method only. The first proposed technique deals with Twitter known terminologies which are RT, Hashtags, Emoticons, and Target; the system consisted of three parts, Pre-processing of Tweets, Scoring Module, and Tweet Sentiment Scoring. The use of corpus technique instead of applying a dictionary such as SentiWordNet led to inconsistency in scores when changing the corpus. While the second approach obtained a medium accuracy because of the manual subjective labelling, it cannot be classified as a real hybrid system because it is actually depending only on the corpus. The third approach reached 80% accuracy ratio which is a little bit better than the first approach but it is also suffering from the corpus-based issue which is obtaining varying results when the corpus is changing.

In [29], Bag of words (2 dictionaries one for positive and one for negative) was used for polarity assignation; the tweet final polarity was calculated by scoring and aggregation; also,

SentiWordNet and Afinn – 111 dictionaries were tested and the results of all of these dictionaries were compared, as a result, SentiWordNet was the best.

A customized dictionary was used in [30] for video textual feedbacks in the Chinese language for analysis purpose; simple computations were used for calculating negative and positive points for each feedback and the result of comparing negative and positive points was the final sentimental result.

The accuracy of using dictionary based approach was shown by [31] , the article show that the accuracy of using WORDNET dictionary is around 98% , which is extremely high in compare with other techniques such as Linguistic Inquiry and Word Count or Context dependent lexicon.

Sentiment analysis based on sentiment dictionary was proposed in [32], the main objective was to help network controllers to achieve effective public opinion administration and make the finest choices. The experimental results indicated that sentiment dictionary method can meritoriously and truthfully analyze micro-blog's sentiment.

2.3.2 Fuzzy Logic approach

Applying Fuzzy Logic in business for reviewing products for opinion mining was presented in [33], authors define six rules for fuzzy logic and the extracted features were filtered by applying a threshold one frequency limit, so if the feature occurrences are below the specified threshold limit then the feature will be ignored; the score of each opinionated word was calculated using the SentiWordNet database, and the results were classified using the SimpleKMeans clustering algorithm. One of the main advantages of this paper is using the Fuzzy Logic on real data that is related to the business field to show the possibility of using the sentimental analysis to track the business and gain benefits.

A discussion about different levels of opinion mining, sentiment lexicon, issues and challenges in Opinion Mining and Sentiment Analysis from the concept of Fuzzy logic was presented in [8], the main objective was using Fuzzy logic for opinion mining problems in Sentiment Analysis. The paper discussed all the above issues in details explaining all the main issues in sentiment analysis, where to use fuzzy logic and what are the limitations for fuzzy logic.

2.3.3 Sentiment analysis on social media

A core NLP system was presented in [34], for this purpose, authors developed a POS tag-set for Twitter and features for Twitter POS tagging were developed on conducted experiments for evaluation. Three manual tagging stages were done and a part-of-speech tagger for Twitter was

developed. The paper contains many implemented features that are not fairly explained and the manual tagging process was not the best choice for such an approach.

The possibility to create an Arabic sentimental analysis tool that handles the dialects and the Arabizzi was the question raised in [35], the main objectives were mapping dialect and Arabizzi to modern standard Arabic. Authors used dialect and Arabizzi dictionaries that were manually collected, they apply there methodology on a small data-set, and they used only the three basic sentiment classifications (positive, negative and neutral), also they relied on an existed solution, all of these factors are limitations for the accuracy of the results in the paper, in addition, that, inaccurate results were detected in the results of the paper (surprising results) and the authors assumed that they happened because of the small size of the data-set.

In [36], in order to extract implicit information from the content of Facebook posts, authors categorized Facebook posts into seven categories and they worked based on the assumption that posts in social network sites have significant information content in addition to the phatic aspect and they worked based on a questionnaire. The study presented a concept but it did not apply it to solve a problem, the authors did not provide results that could be used in a system directly and they only show that posts have implicit information that can't be detected by applying the well-known sentimental analysis approaches.

[37] Presented quantification for the sentiments by using the emotion (happy, unhappy, and emotionless) patterns on Facebook. A randomly selected post from the university's (University Teknologi MARA, Malaysia) Student Financial Section (Bahagian Pengurusan Kewangan Pelajar UiTM) Facebook page in Malay and English languages were collected and Sentiment analysis was done by using Sentiment Identification Algorithm which is Compositional Semantic Rule, Numeric Sentiment Identification, and Bag-of-Word and Rule-based, also emoticons were also classified manually into two categories which are Happy and Unhappy. The result of this paper shows a high emotionless percentage because of taking a sample from one page that belongs to finance students, which led to a very low benefit from the suggested system; authors should use a different resource of data to have more results that can be classified to positive and negative.

The objective of [38] was sentimental analysis for tweets in public Arabic, the main hypotheses were applying supervised sentimental analysis on Arabic language and testing the benefit of translating Arabic words to English words to obtain the English dictionaries benefits. The paper took the benefit of emotions and Google Translate; according to native Arabic speakers

review, the accuracy of using Google translate reached 88% and that result was in 2016, which is heavily reflected the usefulness of using Google translate in Arabic sentiment analysis.

Turkish sentiment analysis lexicon were developed in [39] to analyze collected data from Twitter, authors developed a comprehensive sentiment lexicon that consist of 5,405 words selected from the words frequently used regularly in Turkish language and were also among the most frequently used words in the tweets from the obtained data.

2.3.4 Lexicon analysis

Finding the relationship between corpuses words was the objective of [40]. LogDice which is a relation between words in the corpus was presented as a new association score. The author started from where others end, almost all association score formulas use frequency characteristics from a contingency table, which records the relationship between two words; the author defined a matrix that summarizes the current notation of frequencies of words, then they summarized formulas of some association scores like T-score, Minimum Sensitivity and Dice coefficient, then they apply them on examples, after that they presented the LogDice formula. The authors took the Dice as a base, values of the Dice score are usually very small numbers, this problem was fixed by LogDice, LogDice score has a reasonable interpretation, scales well on a different corpus size, and it is stable on subcorpora, and the values were in a reasonable range.

The objective of [41] was the same objective of [40]; Refined Dice was presented as a new relation between words in a corpus. Refined Dice also took the Dice as a base and built over it because satisfying the performance in extracting significant collocates and co-occurrences.

2.3.5 Pros and Cons

Based on the literature review, a set of articles were selected to compare them with the proposed systems. The comparison result is summarized in the table below:

Paper	Targeting Classical Arabic?	Targeting Social Media?	New model ?	Number of classifications	Sample Size	Continues Learning?	Fully Automated
Duwairi & Qarqaz, 2014	√	√	✗	3	Big	✗	✗
Hardeniya & Borikar	✗	✗	√	3	Big	✗	√
Ahmad & Rahmath, 2014	✗	✗	√	3	Big	✗	√
Duwairi , Marji, Sha'ban, & Rushaidat , 2014	✗	√	✗	3	Small	✗	√
Proposed System	√	√	√	7	Big	√	√

Table 2-1Pros & Cons

2.4 Summary

In this chapter, we discussed the desirable background topics and the previous researches to discover the gaps in Arabic sentiment analysis on social media to work on them in our research. We categorize the studied algorithms according to their approaches.

According to our review, we choose to work on a hybrid system based on the dictionary and the fuzzy logic with applying emoticons detection and Google translate to obtain the best possible accuracy in sentimental analysis with limited employment for the ontological rules of the Arabic language.

In the next chapter, we will highlight the used methodology and we will discuss the architecture of the proposed solution in details.

Chapter 3: Data and Methods

3 Data and Methods

The proposed model represents a novel enhancement in terms of technology, techniques, and modelling in compare with state-of-art systems according to the literature review. The proposed model is presenting a new method in applying sentiment analysis on data collected from social media and it is building over the state-of-heart solutions that were proposed before in the field of sentiment analysis.

3.1 Introduction

Classical Arabic Analysis System (CAAS) is a hybrid system composed of dictionary-based approach and fuzzy logic for classifying Classical Arabic sentences based on their sentimental values. In this chapter, the methodology and the implementation of CAAS were highlighted and discussed in details.

CAAS comprised of two main modules which are labeling module for extracting the sentences' characteristics, and classifier module for obtaining the classifications of the sentences.

Dictionary was used for obtaining word-level polarities; obtaining Classical Arabic words polarities is one of the challenges that CAAS is targeting. Fuzzy logic was used to obtain the final sentences' classification using three factors which are Sentence Polarity using the dictionary, Sentence Likes' Ratio obtained from labeling module, and Account's Orientation which is also obtained from labeling module, to infer the classification of the sentence.

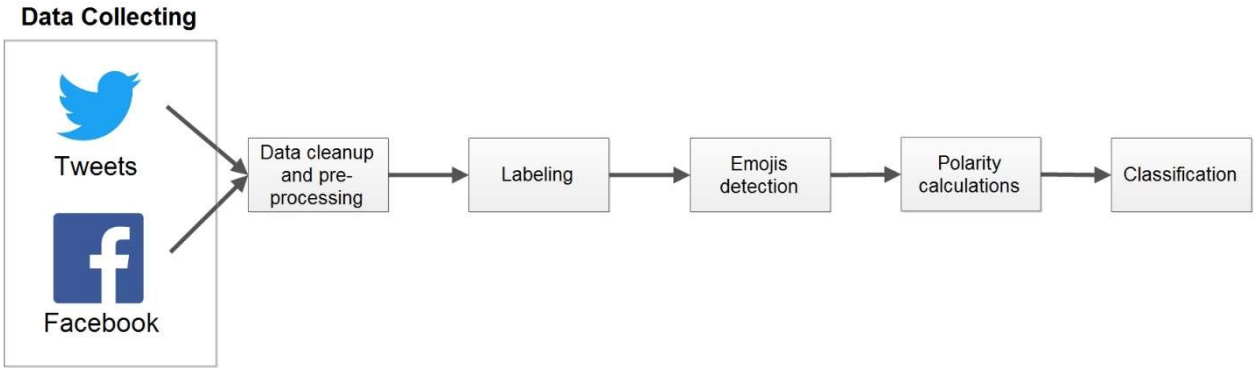

Figure 3-1 Basic Structure of CAAS

CAAS flow starts with obtaining sentences from Twitter and Facebook, the data forwarded directly to the cleanup phase were the sentences become clean and ready for processing, after that, labeling is done to obtain likes' ratio and accounts' orientation, at the same time, the sentence's

polarity is calculated, finally the classifier classifies the result based on the fuzzy rules applied by using the three fuzzy factors and the results will be normalized to have a standardized results.

3.2 Proposed Methodology

The objective of the proposed system is to classify the sentences by measuring their sentimental values. CAAS is represented by two main modules: the labeling module and classification module. The labeling module is responsible for textual analysis for the sentences to obtain the features of them and mathematical calculations based on the obtained data to calculate some other features. Then, the classification module is applied using the fuzzy logic algorithm with the dictionary to obtain the classification results.

The methodology of the proposed model is working as shown in the diagram below:

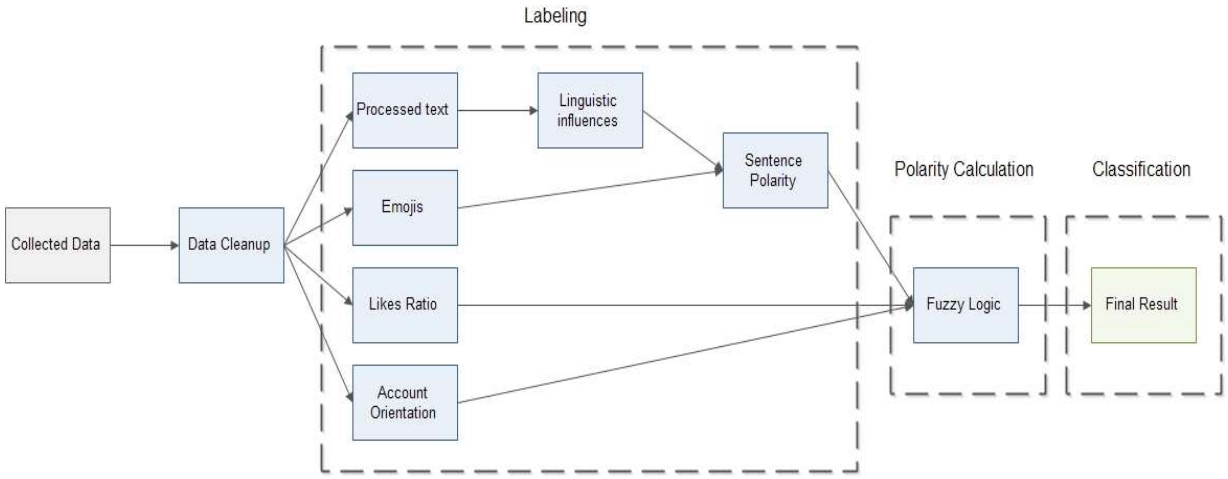

Figure 3-2 CAAS High Level Methodology

3.2.1 Data Collecting

The data collecting phase is the first phase that is responsible for gathering the sentences from the social media pages. In the data collecting phase, the sentences were gathered from Facebook and Twitter then they were stored as row data in the database.

3.2.2 Pre-Processing (Data Cleanup)

The pre-processing phase is a key and essential phase that is responsible for reshaping the collected sentences in an acceptable format for labeling. All the noises such as URLs, stop words,

punctuations, non-Arabic words, and unneeded symbols are removed to make the sentences feasible for the labeling process.

Pre-processing phase result ignores some sentences that are not valid like non-Arabic sentences or sentences that become empty after removing all the noise from them. Also, results of the pre-processing are stored side by side with their original sentences to obtain an original-processed comparison.

The following table is showing some examples of how the pre-processing phase works:

Sentence	Processed Sentence
الضـغوط السـياسـية تتصـاعد على #تيريزا_ماي للتنحي من منصـبها مقابل البحث عن مخرج لأزمة الخروج من الاتحاد الأوروبي	الضـغوط السـياسـية تتصـاعد للتنحي منصـبها مقابل مخرج لأزمة الخروج الاتحاد الأوروبي
كما يطالب الشارع العراقي".. مجلس النواب يصوت على " إقالة محافظ #نينوى ونائبيه بعد كارثة غرق العبارة في نهر #دجلة بالموصل	يطالب الشـارع العراقي مجلس النواب يصـوت إقالة محافظ ونائبيه كارثة غرق في نهر

Table 3-1 Pre-Processing examples

3.2.3 Labeling

Labeling module is used to extract the sentences' dependent features from the textual content of the sentences. Textual analysis tools were built inside CAAS for labeling the sentences. The following steps were done to perform labeling for each triggered account:

1. Reading the processed sentence text and tokenize its textual content to terms using a tokenization tool built in the system.
2. Detect the Facebook and Twitter terms like tweeter or poster name and remove them from the text to ignore them from the analysis process.
3. Detect the emoticons and assign polarities to them based on a defined table that is assigning polarity for each emoticon based on the textual description of each emoticon.
4. Detect negation words, a list of negation words is checked in each sentence and the affection of them on the polarity is taken into consideration rather than it is a shifting for the negative side or the positive side.

5. Calculating the Account orientation indicator which is an indicator of the account that the sentence was collected from, the indicator is calculated based on the following flow:
    a. The range is divided into three regions.
    b. If the positive or the negative exceed 50% of the total amount of the sentences, then indicator value is decided by a fuzzification method.
    c. If the positive and the negative are equal to each other, then the indicator will return zero which means that the indicator is neutral.
    d. If the positive and the negative did not reach 50% of the total amount of sentence, then the indicator is neutral.

    Account orientation indicator is reflecting the overall or the general orientation for an account until the sentence collecting time, so it is a changeable value that is affecting the polarity of the sentence based on the time it was collected in.
6. Calculating the Likes Ratio indicator which is an indicator of the sentence popularity with respect to the popularity of its account, the indicator is calculated based on the following flow:
    a. The range is divided into three regions.
    b. The ratio of the current sentence likes to the max sentence's account likes is calculated.
    c. The indicator value is decided by a fuzzification method.

3.2.3.1 Using Dictionary

The approach of the proposed system is a hybrid approach consists of the dictionary and fuzzy logic. The dictionary phase is consisting of the use of one dictionary which is SentiWordNet to obtain the polarity of each word in the sentence. To find the polarity of an Arabic word the following approach was used:

1. Translate the word using Google Translate.
2. Check the polarity in SentiWordNet.
3. If the word does not exist in the SentiWordNet, then get the root of the word.
4. Check the root polarity in SentiWordNet.
5. If the root does not exist in SentiWordNet, then categorize it as Neutral.

This approach was used in [38] and it reached an accuracy of 88%. At the end of the dictionary phase, a word-level polarity is calculated and it is used as an initial polarity for the sentence.

Dictionary phase was done to obtain word-level polarities which were used to produce sentence level polarities to determine one of the Fuzzy Logic inputs, which is the Sentence Polarity Classification.

3.2.4 Sentence Polarities calculation

After the labeling process, the polarity of a sentence is calculated using the following formula:

$$\text{Sentence Polarity} = \frac{\sum \text{Sentence Term Polarity} + \sum \text{Emoticon Polarity}}{\text{Terms Count} + \text{Emoji Polarity Count}} \quad (1)$$

Sentence polarity was calculated and fuzzified to be used as one of the factors in determining the classification of the fuzzy logic.

3.2.5 Sentence creation time effect

The sentence creation time effect is the affection of the displacement of the sentence creation time since the last sentence in the same account to the average of the creation time for all of sentence of the same page. Formula 2 is explaining the sentence creation time effect.

$$Sentence\ Polarity\ (SP) = \begin{cases} SCT < Average\ SCT, & SP = SP + Average\ SCT/SCT \\ SCT > Average\ SCT, & SP = SP - SCT/Average\ SCT \\ SCT = Average\ SCT, & SP = SP \end{cases} \quad (2)$$

3.2.6 Classification

After the polarity, account orientation and likes ratio values were calculated, the Fuzzy Logic Rules were applied to classify polarities into seven categories that are dividing the classification range to Very Positive, Positive, Good, Neutral, Not Good, Negative, and Very Negative.

Based on the literature review, no one before used more than five categories like [30] and [42]. CAAS used seven categories which helped in generating more specified and realistic results.

3.3 Classical Arabic Analysis System (CAAS)

CAAS is a sentiment analysis web-based system that is working based on the methodology explained in section 3.2, it aims to analyze accounts' sentences (posts and tweets) sentimentally using Facebook and Twitter as data sources.

3.3.1 Architecture of CAAS

CAAS processes were divided into six main phases, each phase is a preparing phase for the next one and they are working in a sequential manner. The phases are:

1. Data Collection Phase: Data collection is the first phase in the system; Facebook and Twitter were selected as data sources. This phase consists of three sub-phases:
    a. Choosing accounts: a set of accounts were selected for testing the system, those accounts will be mentioned later in the 3.3.2.2 section.
    b. Collect sentences: sentences were collected using LinqToTwitter and custom Facebook parser by tracking the account's name in a real-time manner.
    c. Storing sentences: obtained sentences were stored in the sentences table to prepare them for the Pre-Processing phase.

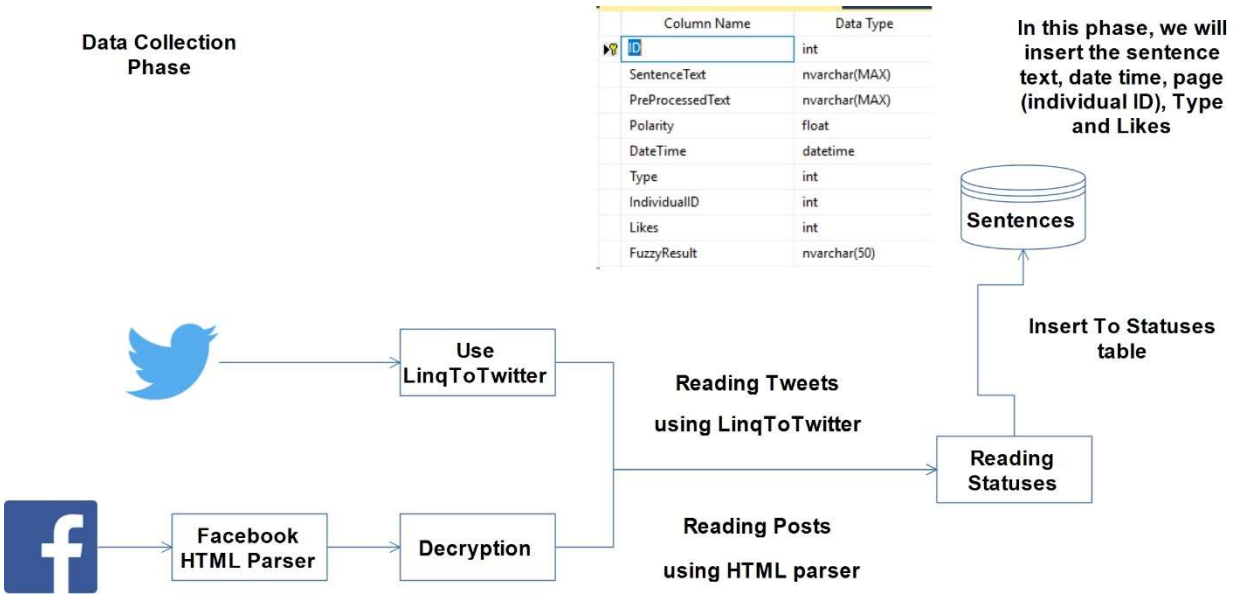

Figure 3-3 Data Collection Phase for CAAS

The Pseudocode for the Data Collection phase is explained in appendix A in 7.1.

2. Pre-Processing Phase (Data Cleanup): Pre-Processing is the second phase in the system; it is an essential phase to transform the raw sentences into an analyzable format. The Pre-Processing phase is consisting of four main steps:
    a. Initial Text Filtering: in this step, the punctuations are removed except the exclamation marks using a filtering tool built inside the system.
    b. Tokenizing: in this step, the sentence's text is divided into tokens; this process includes removing duplicated spaces to normalize the text and preparing it for next processes using a tokenization tool built inside the system.
    c. Removing StopWords: StopWords are sentiment-less words, Ranks NL StopWords list were used as StopWords resource; in this step, StopWords were removed from the sentence's text.
    d. Advance Text Filtering: this step includes different processes that are generating an analyzable text; the processes are removing URLs, removing numbers, removing Twitter symbols, removing Facebook symbols, and remove non-Arabic letters.

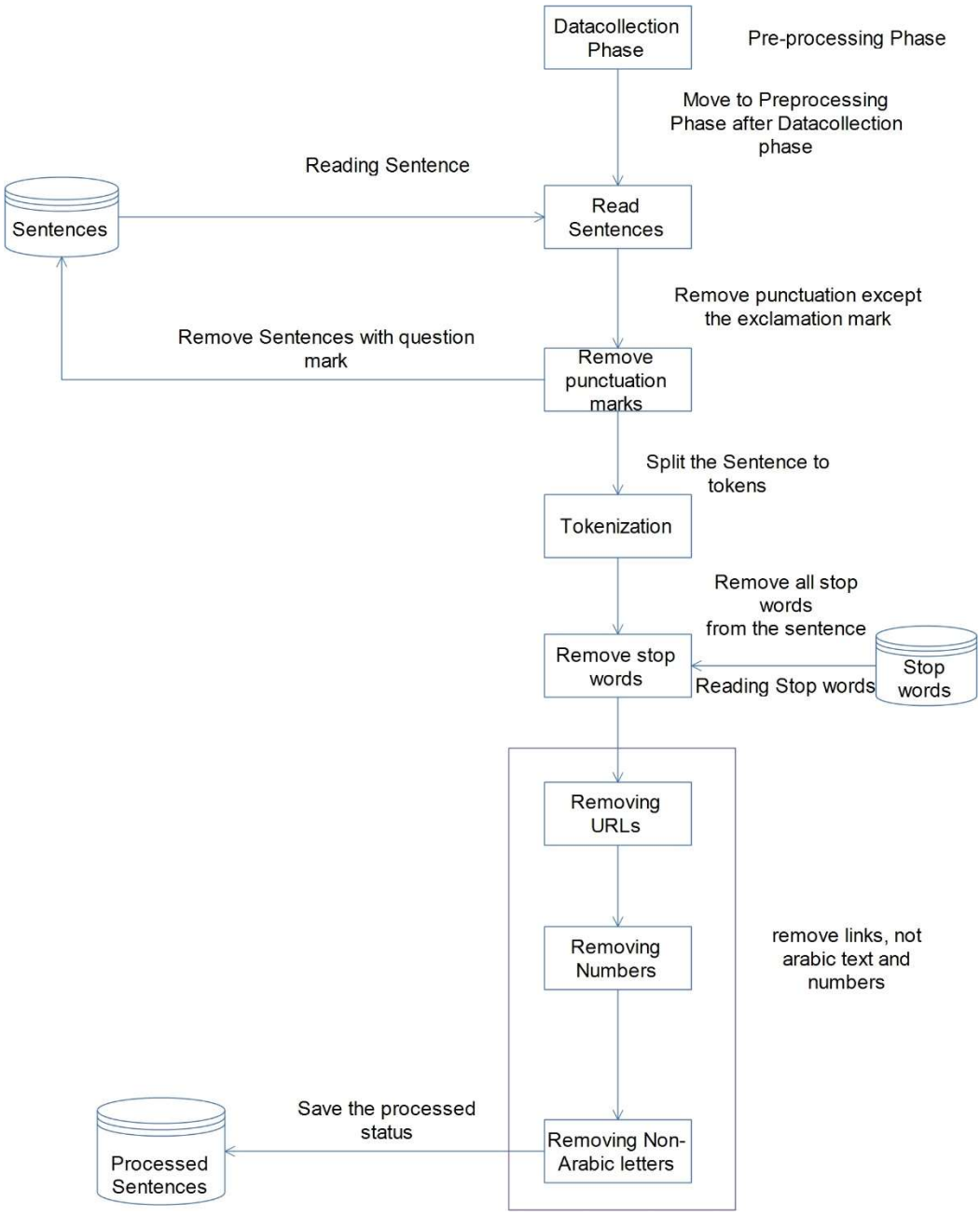

Figure 3-4 Pre-Processing Phase for CAAS

The Pseudocode for the Pre-Processing phase is explained in appendix A in 7.2.

3. Labeling Phase: in this phase, multiple sub-phases were applied to calculate the Sentence Polarity Likes Ratio, and the Account Orientation.

a. Emojis Detection: in this phase, a dictionary of emojis were used to obtain the polarity represented by each emoji included in the sentence text. Emojis polarities were used with the word-level polarities to generate the sentence polarity. Emoticons were manually collected from (emojipedia.org), the polarity for each emoticon was calculated using SentiWordNet by aggregating the polarities for the emoticon description terms.

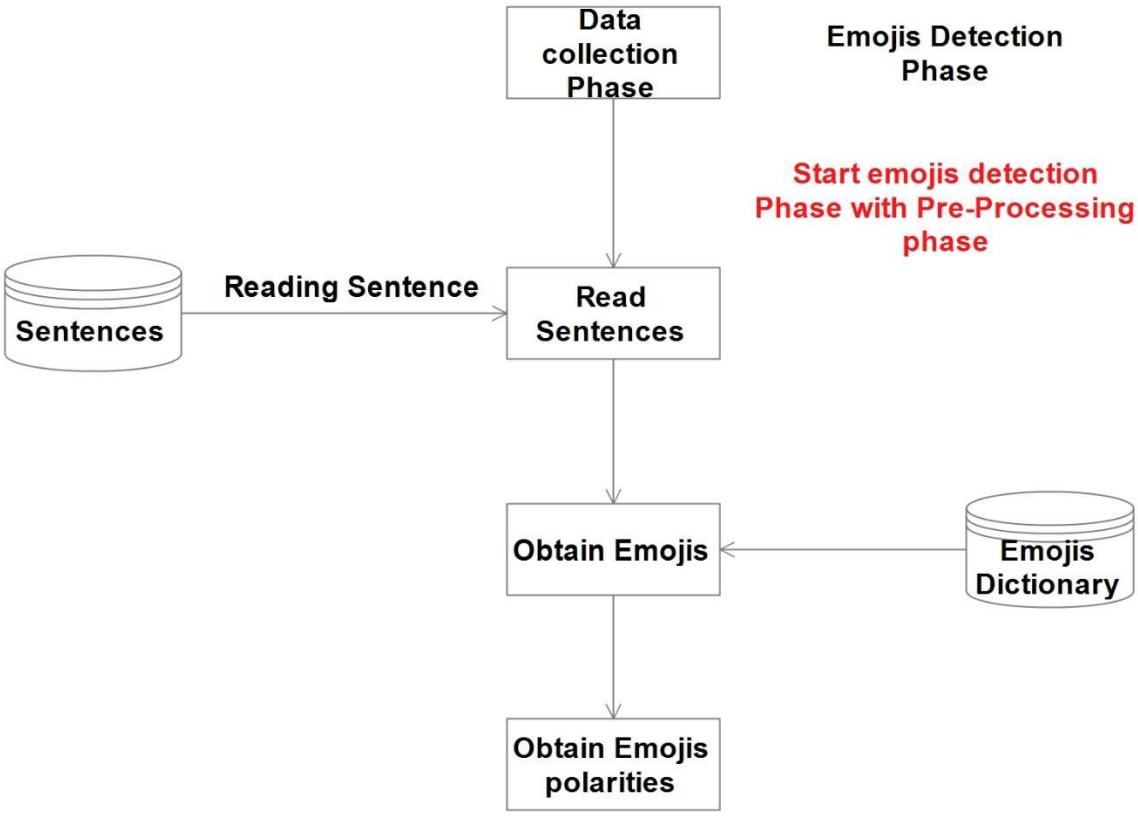

Figure 3-5 Emojis Detection for CAAS

The Pseudocode for the Emojis Detection phase is explained in appendix A in 7.3.

b. Sentence Text Polarity: in this phase, SentiWordNet was used to obtain word-level and sentence level dictionary-based polarities. While SentiWordNet is for English terms, Google translate was used to obtain the English translation by a custom HTML parser built for this purpose.

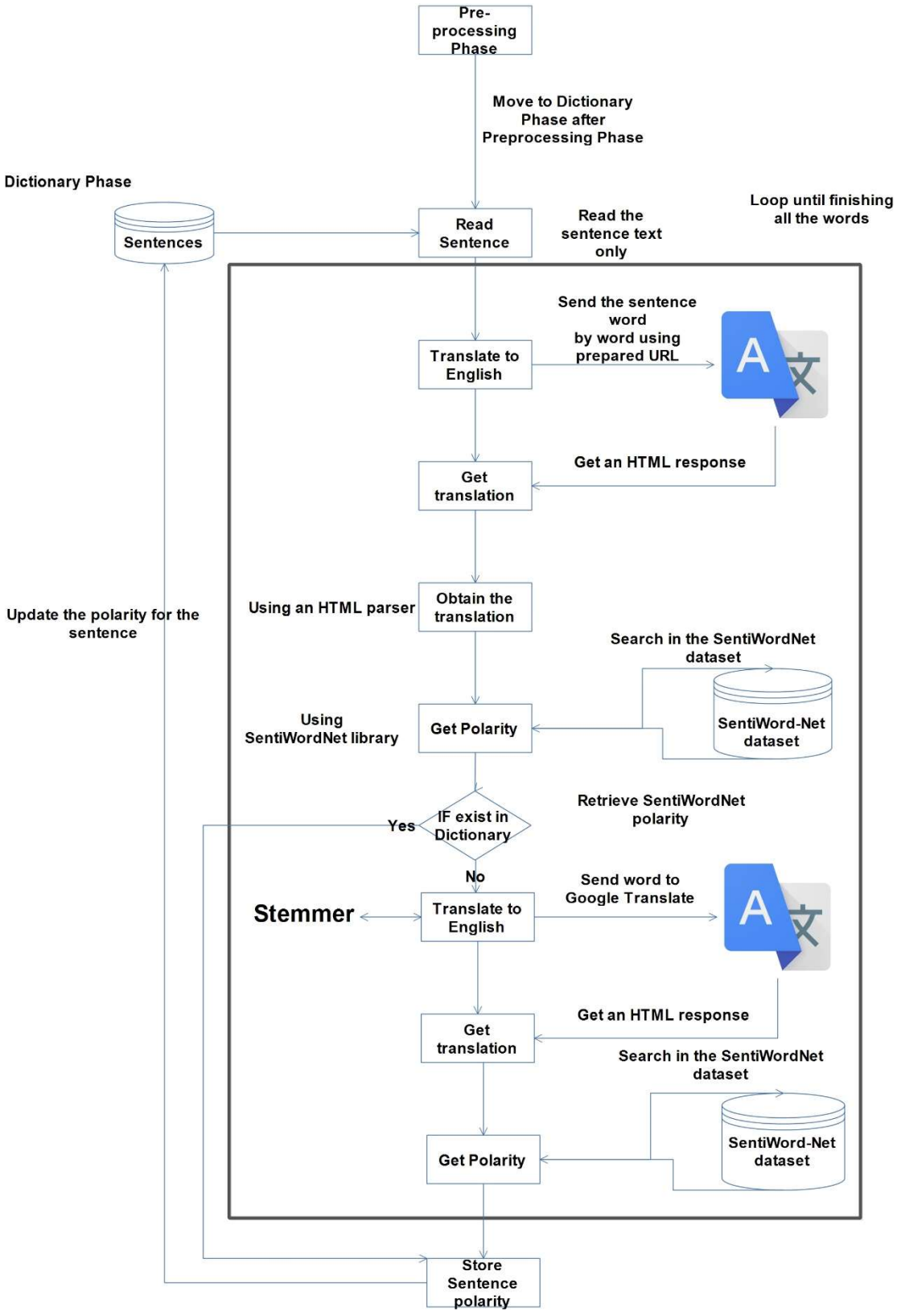

Figure 3-6 Dictionary Phase for CAAS

c. Sentence Polarity: in this phase, the results from Sentence Text Polarity and the Emojis detection were combined together to obtain the final Sentence Polarity as explained in formula 1.
d. Likes Ratio: in this phase, a simple Likes Ratio is calculated based on the sentence likes and the page likes as explained in formula 6.
e. Account Orientation: in this phase, a simple Account Orientation is calculated based on the page pre-classified sentences as explained in formula 7.

The Pseudocode for the Dictionary phase is explained in the appendix A in 8.4.

4. Fuzzy Logic Phase: in this phase, Fuzzy Logic rules were applied and the results were stored to be used in the Reporting phase.

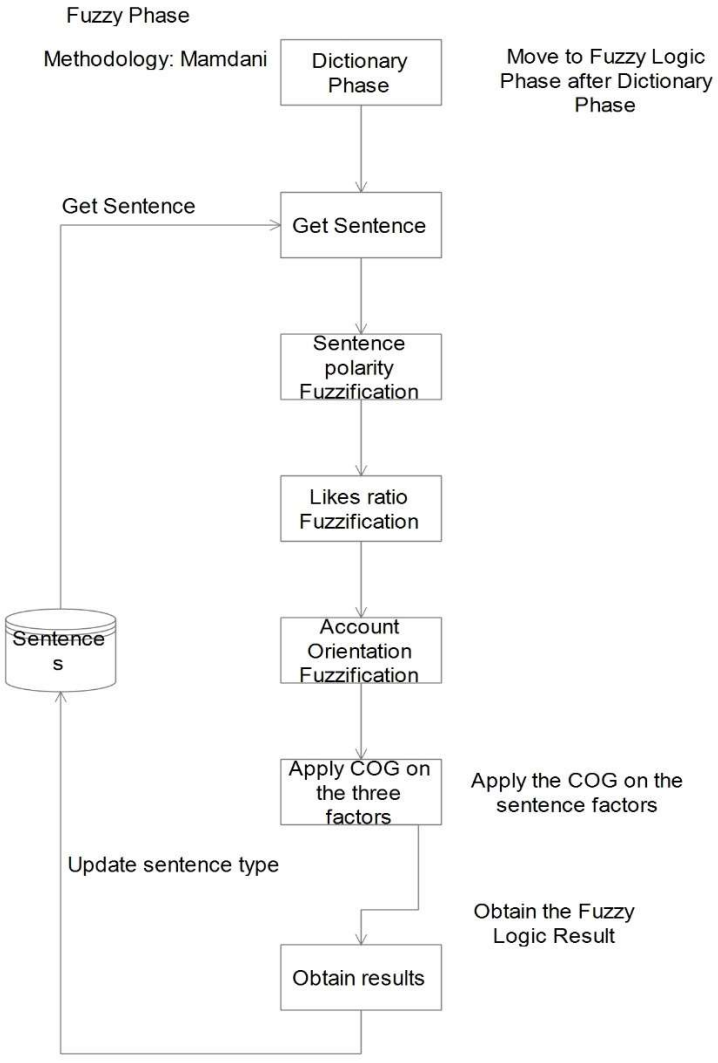

Figure 3-7 Fuzzy Logic Phase for CAAS

The Pseudocode for the Fuzzy Logic phase is explained in the appendix A in 8.5.

5. Reports and Results: the results were presented in a graphical way using charts. The flow of phases and steps above were explained in the following architectural diagram:

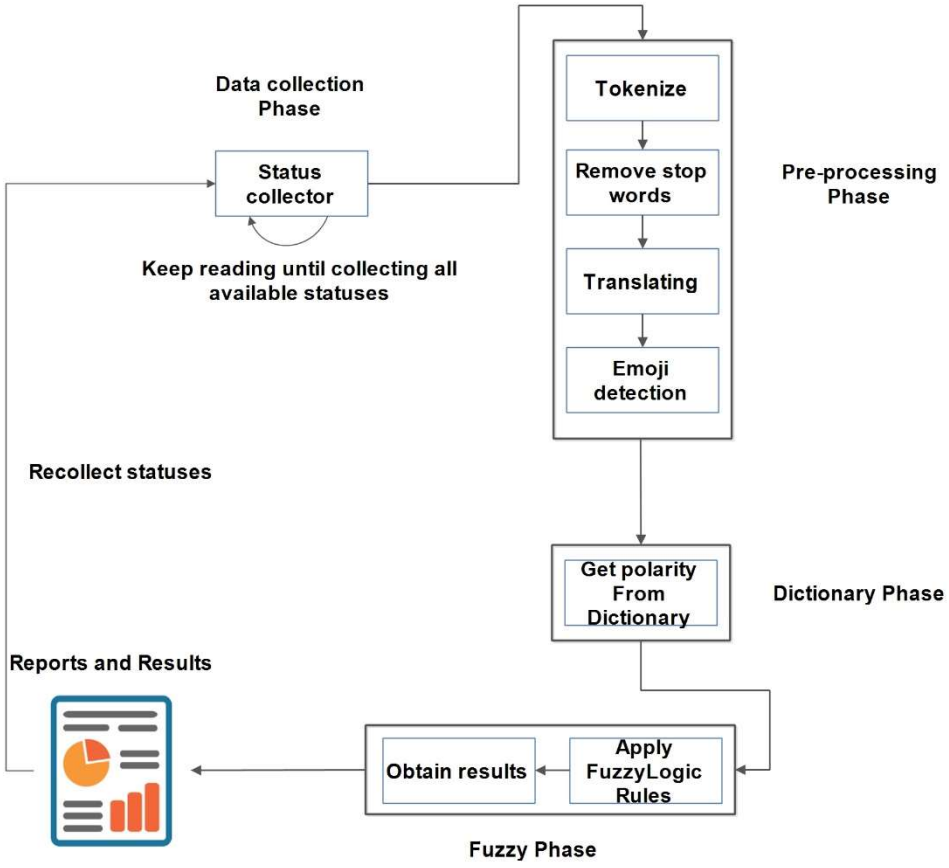

Figure 3-8 CAAS Flow Architecture

3.3.2 Implementation of CAAS

CAAS was implemented using ASP.Net and based on the MVC model; different libraries were used to obtain the maximum possible performance and reliability in the system like LinqToTwitter, LinqToSQL and Entity Framework.

Agile methodology was used in the implementation of CAAS, the agile allow the feedback from each phase to the previous one to solve any unexpected issue. Based on the agile methodology, the system was implemented in the following phases:

3.3.2.1 Database

A database of sentences was used to test the system classification. Different samples with varying sizes were used; the sentences were automatically by CAAS. Each collected sentence was inserted directly into the database in order to move it forward to the pre-processing phase.

3.3.2.2 Data Collection Phase

In this phase, Twitter and Facebook pages were tracked to get sentences. The following table is presenting the official pages on Twitter and Facebook that were tracked for collecting sentences:

No	Name	Official Page	Source
1.	العربية	AlArabiya	Twitter
2.	العربية	AlArabiya	Facebook
3.	الجزيرة	AJArabic	Twitter
4.	الجزيرة	aljazeerachannel	Facebook
5.	RT Arabic	RTarabic	Twitter
6.	RT Arabic	rtarabic.ru	Facebook
7.	Sky News	skynewsarabia	Twitter
8.	Sky News	SkyNewsArabia	Facebook
9.	وكالة معا	MaanNewsArabic	Twitter
10.	وكالة معا	MaanNews.net	Facebook
11.	الملكة رانيا	QueenRania	Twitter
12.	الملكة رانيا	QueenRania	Facebook
13.	أحمد الشقيري	shugairi	Twitter
14.	أحمد الشقيري	AhmadAlShugairi	Facebook

Table 3-2 Data Source Accounts for Data Collecting

The set of accounts were selected taken into consideration that those accounts are usually posting or tweeting in Classical Arabic, the list included news agencies and famous people.

3.3.2.2.1 Collecting from Twitter

LinqToTwitter is an open source third-party LINQ Provider supported and developed by Twitter for the micro-blogging service on Twitter. LinqToTwitter uses standard LINQ syntax for querying, and it includes methods for programmatically using Twitter via APIs.

LinqToTwitter allows the system to get the tweets from the selected Twitter account, with a maximum limit of 200 tweets each time, also it allows the system to get the number of likes on each tweet, which are the needed parameters to know about each tweet (text and likes).

3.3.2.2.2 Collecting from Facebook

Custom HTML parser was implemented inside CAAS to obtain posts from Facebook pages, the parser downloads the page textual content in HTML format with UTF8 encoding, and then the text tokenizer handles the task of splitting the page text to a list of posts.

The obtained posts were encoded in Unicode, a special regular expression was used to convert from Unicode to Arabic. The second phase was to obtain the number of likes for each post, a special multilevel JSON parser was implemented for this purpose. The final result was having a list of pairs for post text and post likes.

3.3.2.3 Pre-Processing Phase

In this phase, several tools were implemented for different purposes, here is a list of those tools:

1. Punctuations detector: Punctuations detector was used to detect and remove any punctuation except the exclamation mark (!).
2. Tokenizer: Tokenizer was used to split the tweet text into words (terms) based on the white spaces.
3. StopWords Detector: used to remove StopWords from the sentence text using the StopWords list, a snap from the list is shown in the table below.

عند
عليها
سنة
احد
بن
به
ثم
ايام

Table 3-3 StopWords Snap

4. URL Detector: used to detect and remove URLs by applying regular expression queries.
5. Numbers Detector: used to detect and remove numbers by applying regular expression queries.
6. Non-Arabic Letters Detector: used to detect and remove any non-Arabic letter by applying regular expression queries.
7. Twitter and Facebook Terms Detector: used to detect and remove special terms, for example:
    a. Mention: a word started with (@) symbol, is used to “tag” somebody in a tweet.
    b. RT: used to refer to a re-tweeted tweet.

The table below is summarizing the changes on the sentence text during the pre-processing phase:

Sentence	Tokenization	Removed Stop Words	Final Text
صور.. ما قصة قلعة القشلة التي بناها الملك عبدالعزيز\nhttp:\/\/ara.tv\/zg9jm",	صور	ما	صور
	ما	التي	قصة
	قصة		قلعة
	قلعة		القشلة
	القشلة		بناها
	التي		الملك
	بناها		عبدالعزيز
	الملك		
	عبد العزيز		

Table 3-4 Pre-processing Snap

3.3.2.4 Ontological processing

Natural language (NL) ontology is a branch of both metaphysics and linguistic semantics. Its purpose is to discover the ontological classes, concepts, and structures that are implied in the use of natural language.

Natural language seems to include its own ontology. That is, there are ontological classes, concepts, and structures that seem to be reflected in the semantics of numerous relevant kinds of natural language languages and structures. [43]

Semantics plays a significant role in Arabic language processing, we can't achieve deep Arabic text processor deprived of adequate information on the implication of the semantic associations between words [44].

The ontology of the language is affecting the meaning of words; some words affect others meanings with positive or negative affections. Researchers did not use any mathematical calculations to measure the affection of negation words on other words in Classical Arabic before, in order to fill this gap some of the negation words were defined, three cases of this affection were taken into considerations, these cases were discussed in [45] but for English language, and CAAS applied them for Arabic language:

1. First cases: if ("لن", "لم", "غير", "ليس", "لا") are existed, then the polarity will be raised to the power (0.5).

   Word Polarity with negation word $= (\text{Polarity})^{0.5}$ (3)

2. Second case: if an adverb that give extreme positivity or negativity is existed like "جدا", "كثيرا", "بشدة", then the polarity will be duplicated.

   Word Polarity with positive adverb $= (\text{Polarity X } 2)^{0.5}$ (4)

3. Third case: if a negation word with defined adverb are exist, each one will have its own effect on the next word polarity, the final word polarity will be:

   Word Polarity negation and adverb $= (\text{First Sub Polarity X Second Sub Polarity})^{0.5}$ (5)

3.3.2.5 Dictionary Phase

In this phase, SentiWordNet dictionary was used to obtain the words polarity; SentiWordNet that gives double polarities values from 1.0 to -1.0 for each word, SentiWordNet values were divided into seven categories to match the proposed categorization, the following table is showing the categorization ranges:

Category	SentiWordNet Polarity
Very positive	Higher than 0.66
Positive	Higher than 0.33 and less than or equal to 0.66
Good	Higher than 0 and less than or equal to 0.33
Neutral	0
Good	Less than 0 and higher than or equal to - 0.33
Negative	Less than -0.33 and higher than or equal to - 0.66
Very negative	Less than - 0.66

Table 3-5 Categories of word-level polarity

If the word does not exist in the dictionary, it is categorized as neutral and the polarity zero is assigned to it. From a technical perspective, SentiWordNet is providing two polarities for each word, one polarity for the positive and another one for the negative. To solve this issue, a SentiWordNet.NET library implemented by Chalmers University of Technology was used to extract the final polarity for each word independently.

While SentiWordNet is for English words, Google translate was used for translating in word level, a custom tool for translation was implemented to communicate with Google translate and obtain the translated word to check the translated word in SentiWordNet.

In addition to the translation tool, a stemming tool was implemented to help in obtaining the polarity. If the translated word did not exist in SentiWordNet, the stemmer gets the word stem in Arabic and the translation tool obtain the stem translation to check it in SentiWordNet, if the stem and the word did not exist in SentiWordNet then zero polarity value was assigned to the word.

3.3.2.6 Fuzzy Logic phase

In this phase, the factors that are used in the Mamdani were calculated and used to obtain the sentence sentiment classification. The calculated factors are:

a. **Sentence polarity classification:** the polarity of the textual content of the sentence after pre-processing in addition to the polarities of the emojis in the sentence was sent to a fuzzification method to obtain the fuzzy value for the sentence polarity.

   The Sentence polarity range is divided into seven sub-ranges which are very positive, positive, good, neutral, not good, negative and very negative.

b. **Account orientation:** the account orientation is a dynamically changeable measurement that is indicating the general orientation for a specific page from a sentimental point of view. The account orientation is divided into three ranges which are positive, neutral, and negative; the value of the indicator is measured based on the previously classified sentences collected for the account that the new sentence is obtained from, a comparison between positive and negative tweets is used to determine the account orientation if one of them exceed 50% of the total number of sentences based on a fuzzification method, otherwise the account orientation will be neutral.

c. **Likes ratio:** likes ratio is a dynamically changeable value that is measuring how popular is a specific sentence by using the number of its likes in a comparison with the maximum number of likes obtained by all sentences for the same account. The ratio is measured by dividing the number of sentences likes on the maximum number of the likes for a sentence belongs to the same account, then a fuzzification method is calculating the fuzzy value of the likes' ratio.

3.3.2.6.1 Fuzzification methods

The database of a rule-based system might hold inaccuracies, which appear in the explanation of the rules specified by the expert. Since such an inference cannot be made by the approaches which use traditional Boolean logic, Zadeh in [46] and Mamdani in [47] proposed an inference rule called "compositional rule of inference". Using this inference rule, numerous approaches to fuzzy reasoning were projected. Zadeh [48] extends the old-style Modus Ponens rule in order to work with fuzzy sets, obtaining the Generalized Modus Ponens (GMP) rule [49]. The main reward of this approach is that it does not need a model of the process.

3.3.2.6.2 Fuzzification

Fuzzification is the Process of converting crisp input into fuzzy input in the form of membership function. The ranges for each input and output were divided into regions based on testing perspective, likes ratio was divided to three regions, account orientation was divided to three regions, sentence polarity was divided to seven regions, and finally the output is divided to seven regions. The following diagrams are illustrating the regions' division:

Likes ratio:

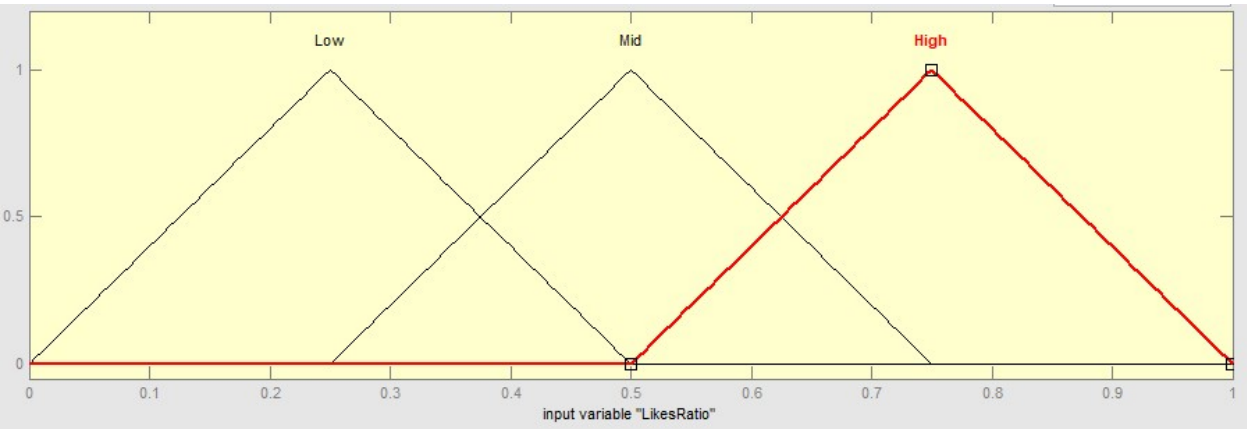

Figure 3-9 Likes' ratio ranges

Likes Ratio is calculated by two steps, in the beginning, a simple ratio is calculated using the following formula:

$$\text{Likes Ratio} = \begin{cases} Max\ Likes = 0, & 1 \\ Max\ Likes = Sentence\ Likes, & 1 \\ Sentence\ Likes = \dfrac{MaxLikes}{2}, & 0 \\ Sentence\ Likes > \dfrac{MaxLikes}{2}, & \dfrac{Sentence\ Likes}{\frac{MaxLikes}{2}} \\ Sentence\ Likes > \dfrac{MaxLikes}{2}, & \dfrac{Sentence\ Likes}{\frac{MaxLikes}{2}} \times -1 \end{cases} \quad (6)$$

After the basic Likes ratio is calculated, Mamdani was used to find the final value; Mamdani is explained in equation 8.

Account orientation

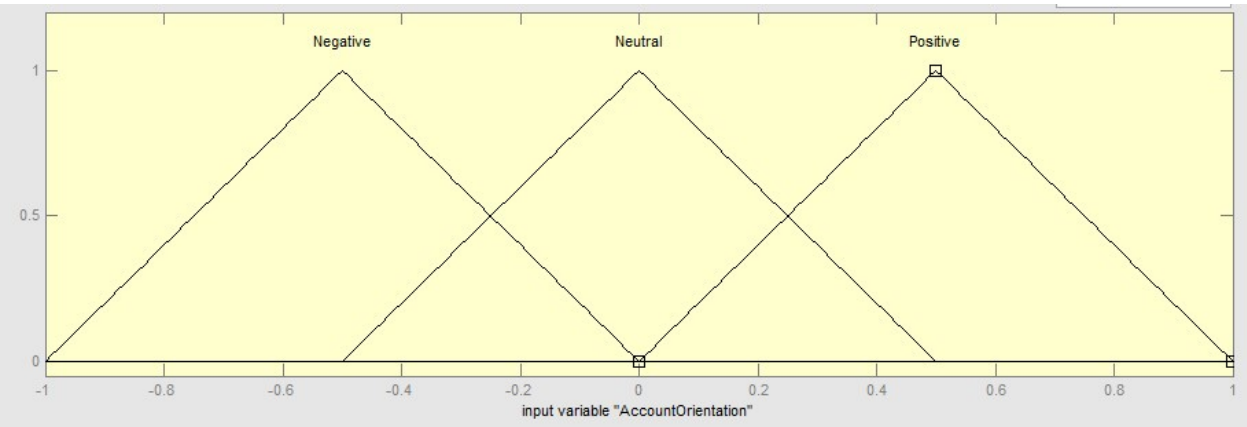

Figure 3-10 Account orientation ranges

Account Orientation is calculated by two steps, at the beginning, simple ratio is calculated using the following formula:

$$\text{Account Orientation} = \begin{cases} Poritive\ sentences\ OR\ Negative\ Sentences < 50\ of\ total\ Sentences, & 0 \\ Positive\ Sentences > Nagative\ Sentences, & \dfrac{Positive\ Count}{Sentences\ Count} \\ Positive\ Sentences < Nagative\ Sentences, & \dfrac{Negative\ Count}{Sentences\ Count} \end{cases} \quad (7)$$

After the basic Account Orientation is calculated, Mamdani was used to find the final value; Mamdani is explained in equation 8.

Sentence polarity

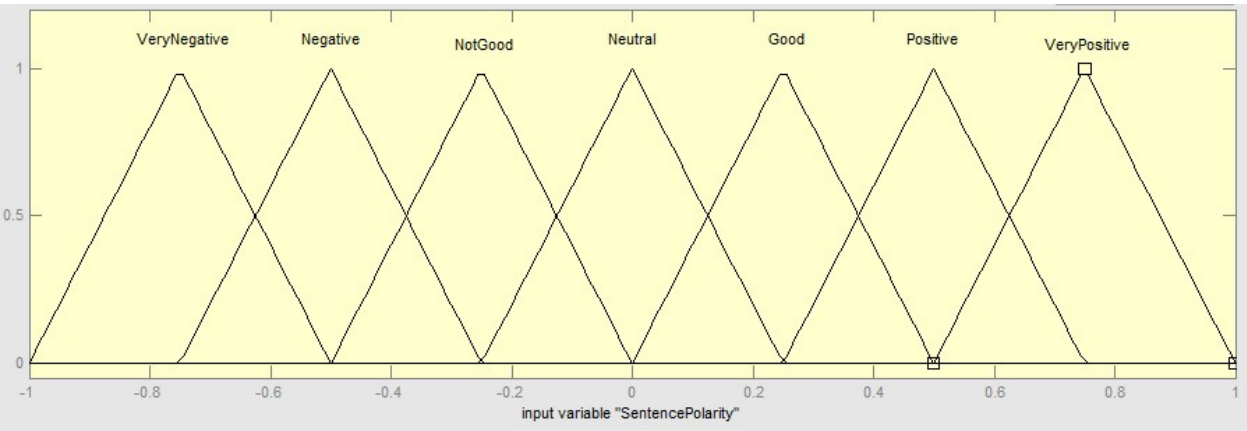

Figure 3-11 Sentence polarity ranges

After the basic Account Orientation is calculated by equation 1, Mamdani was used to find the final value; Mamdani is explained in equation 8.

Fuzzy Result

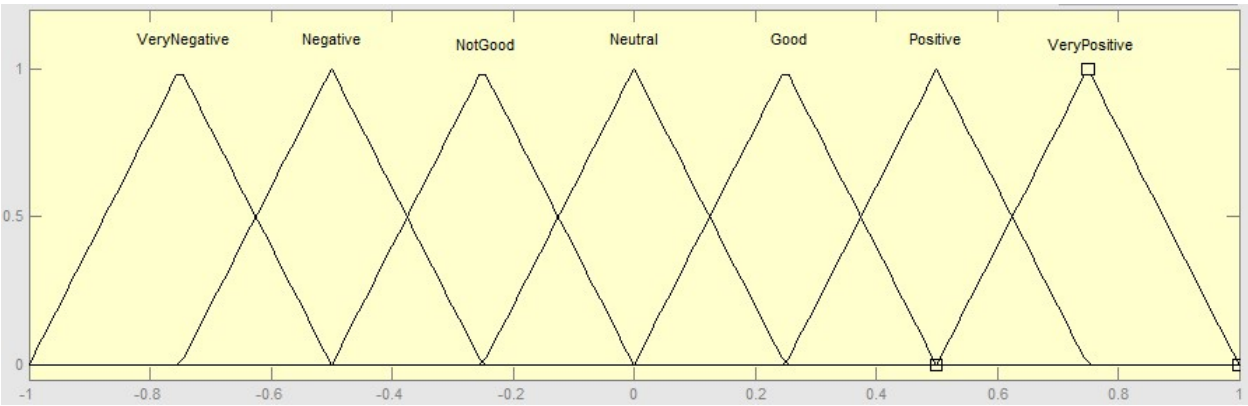

Figure 3-12 Result ranges

3.3.2.6.3 Rule Evaluation

In the Rule Evaluation, a set of rules were defined to determine what action to make in response to inputs.

The set of rules includes 63 rules that combines the Sentence Polarity, Likes Ratio, and Account orientation using AND operator.

1. If (LikesRatio is Low) and (AccountOrientation is Negative) and (SentencePolarity is VeryNegative) then (SentenceClassificatio is VeryNegative) (1)
2. If (LikesRatio is Low) and (AccountOrientation is Negative) and (SentencePolarity is Negative) then (SentenceClassificatio is VeryNegative) (1)
3. If (LikesRatio is Low) and (AccountOrientation is Negative) and (SentencePolarity is NotGood) then (SentenceClassificatio is Negative) (1)
4. If (LikesRatio is Low) and (AccountOrientation is Negative) and (SentencePolarity is Neutral) then (SentenceClassificatio is Negative) (1)
5. If (LikesRatio is Low) and (AccountOrientation is Negative) and (SentencePolarity is Good) then (SentenceClassificatio is NotGood) (1)
6. If (LikesRatio is Low) and (AccountOrientation is Negative) and (SentencePolarity is Positive) then (SentenceClassificatio is NotGood) (1)
7. If (LikesRatio is Low) and (AccountOrientation is Negative) and (SentencePolarity is VeryPositive) then (SentenceClassificatio is Neutral) (1)
8. If (LikesRatio is Low) and (AccountOrientation is Neutral) and (SentencePolarity is VeryNegative) then (SentenceClassificatio is VeryNegative) (1)
9. If (LikesRatio is Low) and (AccountOrientation is Neutral) and (SentencePolarity is Negative) then (SentenceClassificatio is Negative) (1)
10. If (LikesRatio is Low) and (AccountOrientation is Neutral) and (SentencePolarity is NotGood) then (SentenceClassificatio is Negative) (1)
11. If (LikesRatio is Low) and (AccountOrientation is Neutral) and (SentencePolarity is Neutral) then (SentenceClassificatio is NotGood) (1)
12. If (LikesRatio is Low) and (AccountOrientation is Neutral) and (SentencePolarity is Good) then (SentenceClassificatio is NotGood) (1)
13. If (LikesRatio is Low) and (AccountOrientation is Neutral) and (SentencePolarity is Positive) then (SentenceClassificatio is Neutral) (1)
14. If (LikesRatio is Low) and (AccountOrientation is Neutral) and (SentencePolarity is VeryPositive) then (SentenceClassificatio is Neutral) (1)
15. If (LikesRatio is Low) and (AccountOrientation is Positive) and (SentencePolarity is VeryNegative) then (SentenceClassificatio is VeryNegative) (1)
16. If (LikesRatio is Low) and (AccountOrientation is Positive) and (SentencePolarity is Negative) then (SentenceClassificatio is Negative) (1)
17. If (LikesRatio is Low) and (AccountOrientation is Positive) and (SentencePolarity is NotGood) then (SentenceClassificatio is NotGood) (1)
18. If (LikesRatio is Low) and (AccountOrientation is Positive) and (SentencePolarity is Neutral) then (SentenceClassificatio is NotGood) (1)
19. If (LikesRatio is Low) and (AccountOrientation is Positive) and (SentencePolarity is Good) then (SentenceClassificatio is Neutral) (1)
20. If (LikesRatio is Low) and (AccountOrientation is Positive) and (SentencePolarity is Positive) then (SentenceClassificatio is Neutral) (1)
21. If (LikesRatio is Low) and (AccountOrientation is Positive) and (SentencePolarity is VeryPositive) then (SentenceClassificatio is Good) (1)
22. If (LikesRatio is Mid) and (AccountOrientation is Negative) and (SentencePolarity is VeryNegative) then (SentenceClassificatio is VeryNegative) (1)
23. If (LikesRatio is Mid) and (AccountOrientation is Negative) and (SentencePolarity is Negative) then (SentenceClassificatio is Negative) (1)
24. If (LikesRatio is Mid) and (AccountOrientation is Negative) and (SentencePolarity is NotGood) then (SentenceClassificatio is NotGood) (1)
25. If (LikesRatio is Mid) and (AccountOrientation is Negative) and (SentencePolarity is Neutral) then (SentenceClassificatio is NotGood) (1)

Figure 3-13 Fuzzy Logic Rules

3.3.2.6.4 Defuzzification

The last stage after rule evaluation is the defuzzification. In defuzzification, all weighty fuzzy outputs will be combined into precise output variables. One normally used defuzzyfication technique is the Center of Gravity (COG) method.

The final result is obtained by Center of Gravity which is calculated by the following formula:

$$\text{Center of Gravity} = \frac{\text{Likes Ratio} + \text{Account Orientation} + \text{Sentence Polaity}}{\text{LR} \times \text{LR intercetions} + \text{AO} \times \text{AO intercetions} + \text{SP} \times \text{SP intercestions}} \quad (8)$$

In order to find the final values, Mamdani was used to find Likes Ratio, Account Orientation and Sentence Polarity, Mamdani formula is as follow:

$$\text{Mamdani} = \begin{cases} \textit{First Membership Value} < \textit{Second Membership Value}, & \textit{First MembershipValue} \\ \textit{First Membership Value} > \textit{Second Membership Value}, & \textit{Second MembershipValue} \end{cases} \quad (9)$$

First and Second memberships are interceptions that are calculated by Pythagorean law and the laws of trigonometry.

3.3.2.7 Results Normalization

There are some cases that could produce an out of range polarity values. In order to overcome with these situations, the polarity is normalized to fit the range using two simple rules,

first one is if the polarity is higher than 1, then the polarity will be 1; the second one is if the polarity is less than -1, then the polarity will be -1.

For example, if the polarity was 1.01 then the normalizer will reduce it to 1.0, also if the polarity was -1.01 then the normalizer will increase it to -1.0 as in formula 9.

$$Sentence\ Polarity = \begin{cases} Sentence\ Polarity < -1, -1 \\ Sentence\ Polarity > 1, 1 \end{cases} \quad (10)$$

3.4 Collected Data

The tables below show a bit of inequality in the number of the collected sentences for the chosen accounts; that indicates a bit of disparity in the interactivity and popularity of the accounts. The tables below show the number of sentences collected for each account on both Facebook and Twitter.

No	Account	Number collected sentences
1	قناة العربية	50
2	قناة الجزيرة	50
3	CNN بالعربية	50
4	RT Arabic	49
5	Sky News	50
6	وكالة معا	50
7	الملكة رانيا	32
8	أحمد الشقيري	46
Total	377	

Table 3-6 The Number of Collected Data – First Round

No	Account	Number collected sentences
1	قناة العربية	50
2	قناة الجزيرة	50
3	CNN بالعربية	50
4	RT Arabic	50
5	Sky News	50
6	وكالة معا	50
7	الملكة رانيا	31
8	أحمد الشقيري	46
Total	377	

Table 3-7 The Number of Collected Data – Second Round

3.5 Summary

In this chapter, we highlighted the proposed methodology in terms of phases, labeling, dictionary usage, calculations and classifications. We also explain deeply CAAS architecture and the execution details of all CAAS phases.

In the next chapter, we will discuss the testing phase and the results obtain by the proposed system, also we will prove the learning process in the proposed system.

Chapter 4: Testing and Results

4 Experiments and Results

This chapter presents the results obtained from the collected data and highlights the learning process in the system, also, the chapter presents the system's scalability explanation, and end up with results discussion.

4.1 System's Scalability

In order to measure the scalability of the system, we did several reading iterations on different data sizes, we measure the time, processor, and memory consumption in each iteration, the table below is summarizing the measurements results.

The results below were gathered using a machine with the following characteristics:

- Model: Dell Latitude E6410.
- CPU: Core i5 1st generation – 4 cores, 2.4 GHz for each core.
- RAM: 4 GB.
- Operating System: Windows 10 Pro 64-bit (10.0 build 17134).

Limit per account	Total sentences	Time MS	CPU %	Memory %
20	155	323819	29.58589	100
40	306	197988	20.18062	100
60	459	214210	33.98223	100
80	614	256033	36.39589	100
100	764	217926	35.28394	100
200	1524	777378	43.64495	100

Table 4-1 Scalability results

4.2 Sentences' Classifications

CAAS classified the sentences to Very Positive, Positive, Good, Neutral, Not Good, Negative, and Very Negative classifications. The following table shows the obtained results by CAAS for each account (in the first and second round):

Company	Very positive	Positive	Good	Neutral	Not Good	Negative	Very Negative
قناة العربية	1%	0%	93%	0%	0%	1%	5%
قناة الجزيرة	5%	1%	66%	2%	0%	10%	16%
CNN بالعربية	1%	0%	89%	0%	0%	5%	4%
RT Arabic	2.02%	0%	63.63%	1.01%	0%	7.07%	26.26%
Sky News	3%	0%	92%	1%	0%	0%	4%
وكالة معا	2%	0%	96%	0%	0%	0%	2%
الملكة رانيا	14.28%	4.76%	58.73%	15.38%	0%	15.87%	6.35%
أحمد الشقيري	2.17%	0%	84.72%	2.17%	0%	6.52%	4.35%

Table 4-2 Overall sentences' classification (initial polarities per account)

The data was collected in two rounds, then each group was rechecked two times, finally, the data were merged together to obtain the final result. The following charts and tables are showing the collected data in each phase independently:

4.2.1 First group – First Round

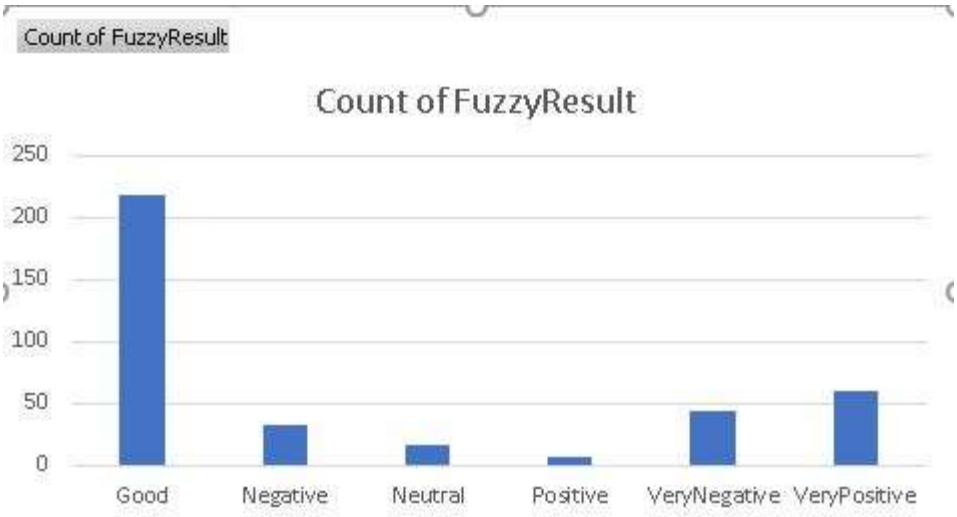

Figure 4-1 First Group - First Round

For the first group, the first round represents the results once the data was collected, the exact numbers for each group are represented in the table below.

Category	Number of sentences
Very Positive	60
Positive	6
Good	219
Neutral	16
Not Good	0
Negative	32
Very Negative	44

Table 4-3 First Group - First Round

4.2.2 First group – Second Round

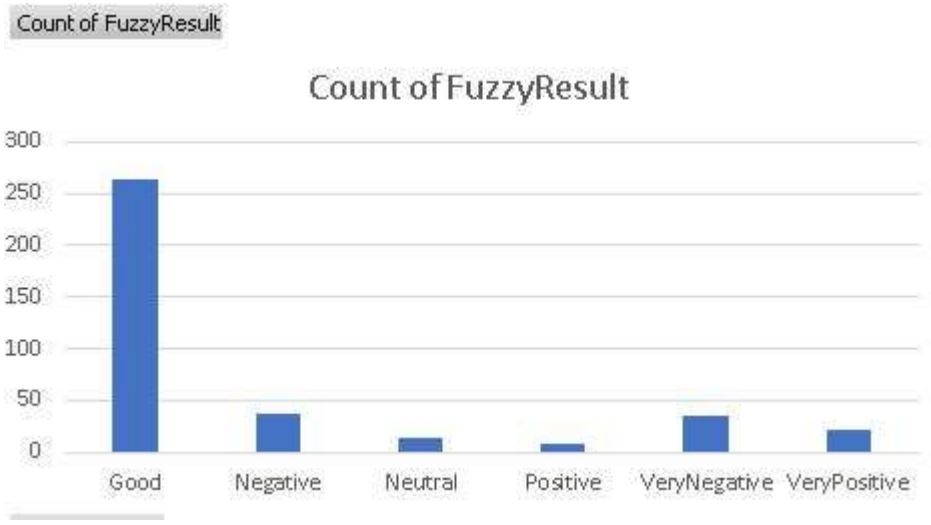

Figure 4-2 First Group - Second Round

The second round was done after the first group was collected, the results were changed because the system recalculated the Account Orientation and the Likes Ratio according to the collected sentences, and the following table is showing the new results.

Category	Number of sentences
Very Positive	21
Positive	8
Good	263
Neutral	13
Not Good	0
Negative	37
Very Negative	35

Table 4-4 First Group - Second Round

4.2.3 Second group – First Round

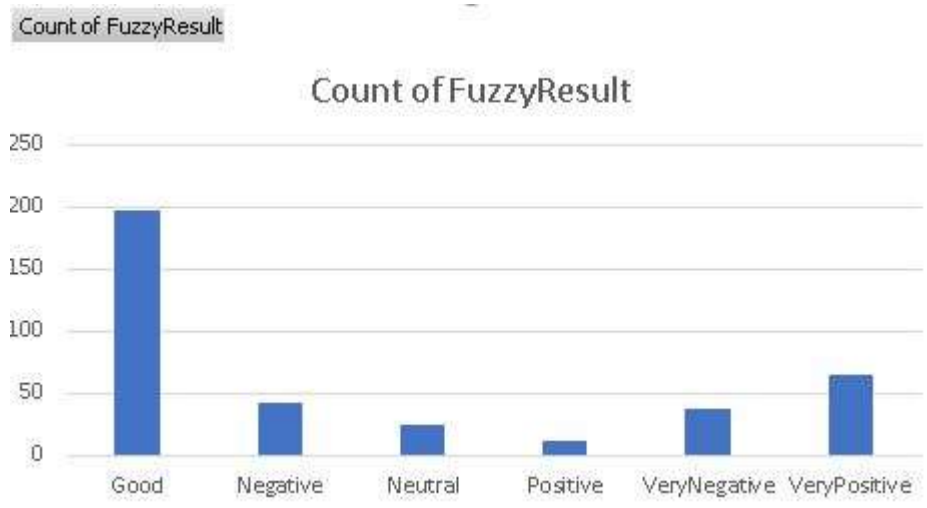

Figure 4-3 Second Group - First Round

For the second group, the first round represents the results once the data was collected, the exact numbers for the sentences in each category are represented in the table below.

Category	Number of sentences
Very Positive	65
Positive	11
Good	197
Neutral	25
Not Good	0
Negative	42
Very Negative	37

Table 4-5 Second Group - First Round

4.2.4 Second group – Second Round

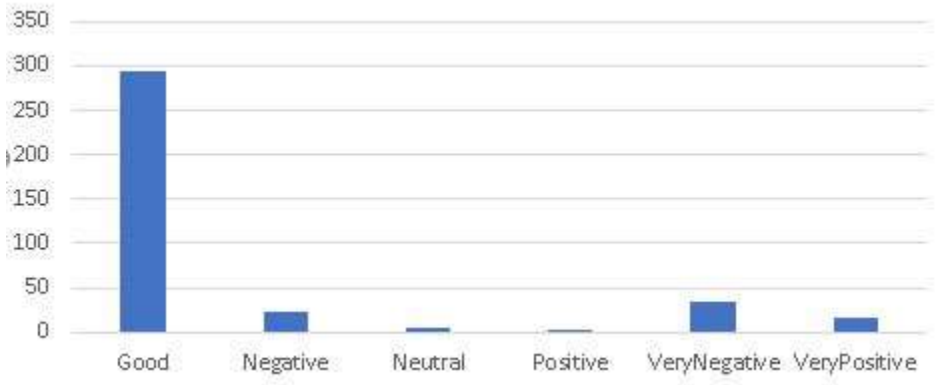

Figure 4-4 Second Group - Second Round

The second round was done after the second group was collected, the results were changed because the system recalculated the Account Orientation and the Likes Ratio according to the collected sentences, and the following table is showing the new results.

Category	Number of sentences
Very Positive	17
Positive	3
Good	294
Neutral	4
Not Good	0
Negative	24
Very Negative	35

Table 4-6 Second Group - Second Round

4.2.5 All Data – Initial categorization

After each group was calculated, we keep a copy from each group to check the changes on the categorization after multiple rounds, the following chart is illustrating all the collected data in the initial round.

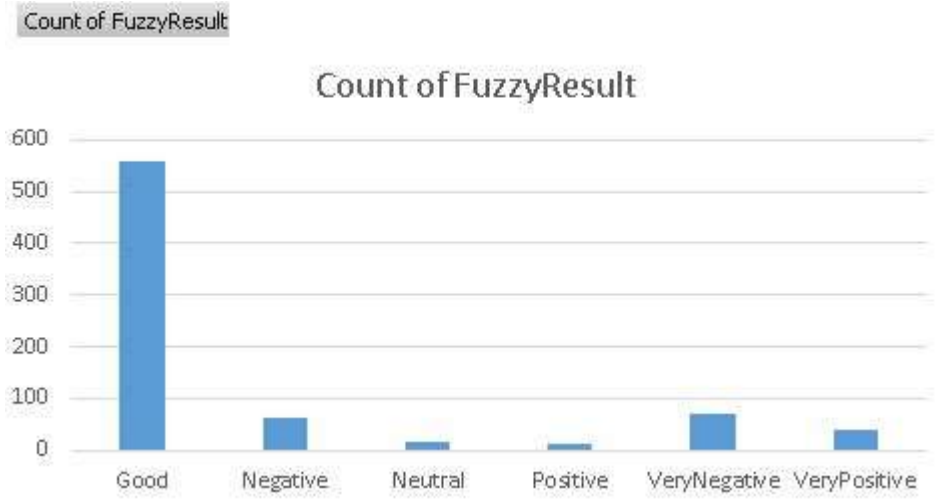

Figure 4-5 All Data – Initial Categorization

The exact numbers for the categories' sentences are represented in the table below:

Category	Number of sentences
Very Positive	38
Positive	11
Good	557
Neutral	17
Not Good	0
Negative	61
Very Negative	70

Table 4-7 All Data - Initial Categorization

4.2.6 All Data – Second Round

After collecting the samples, they were merged together to check categorization changes; the chart below illustrates the categorization changes after one round.

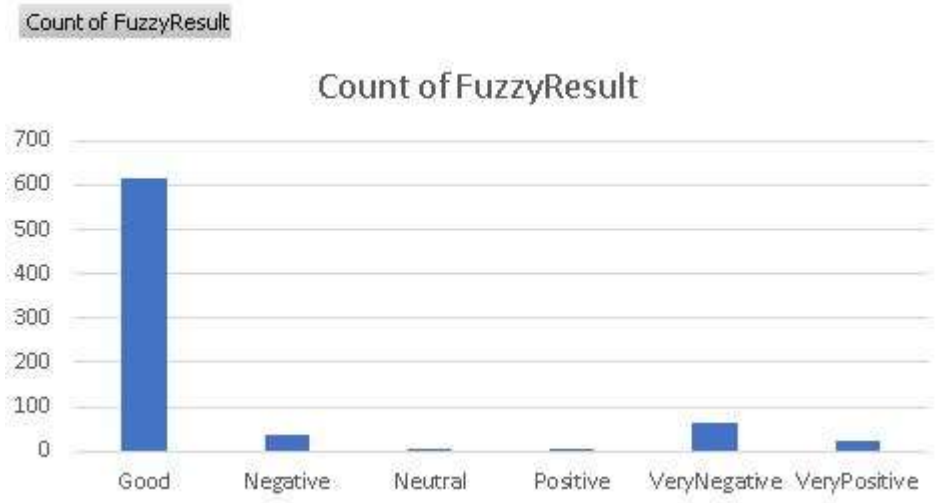

Figure 4-6 All Data - Second Round

The exact numbers for the categories' sentences are represented in the table below:

Category	Number of sentences
Very Positive	25
Positive	4
Good	614
Neutral	7
Not Good	0
Negative	39
Very Negative	65

Table 4-8 All Data – Second Round

4.2.7 All Data – Comparison

Group	Very Positive	Positive	Good	Neutral	Not Good	Negative	Very Negative	Total
First Group - First round	60	6	219	16	0	32	44	377
First Group - second round	21	8	263	13	0	37	35	377
Second Group - First round	65	11	197	25	0	42	37	377
Second Group - Second round	17	3	294	4	0	24	35	377
All Data	38	11	557	17	0	61	70	754
All Data - First Round	25	4	614	7	0	39	65	754

Table 4-9 All Data -Results' Comparison

The table above is showing how the categorization results are changing when the data is changed and accumulate, the groups' results were completely changed and show new categorization for the same data categorized before.

4.3 Mathematical Proof Explanation

The sentences classification process accuracy increased when the system obtained more sentences for the same account; to prove this assumption, we tracked a sentence classification evaluation each time the system gains new data by doing re-classification for the sample with each new input.

The experiment started with one sentence which is:

Note: Average creation time is 720000 milliseconds.

Sentence Text	Processed Text	Polarity	Page	Likes	Fuzzy Result	Creation Time ms
# الكونغرس116 يتصدر قائمة التداول الأمريكية في #تويتر بعد تولّي الديمقراطية نانسي بيلوسي رئاسة مجلس النواب 7	يتصدر قائمة التداول الامريكية تولي الديمقراطية نانسي بيلوسي رئاسة مجلس النواب	1	الجزيرة	25	Very Positive	720000

Table 4-10 Tracking continuous learning - 1 sentence

Mathematical explanation:

1. By applying Equation 1:
   $Sentence\ Polarity = \frac{0.625+0+0+0+0+0+0+0+0+0+0 \quad .166666666666667+0+0}{11} =$
   0.071969696969697
2. By applying equation 2, the creation time is equal to the average creation time, so it has no effect.
3. By applying Mamdani for Sentence Polarity (Equation 9):
   First Membership value= 0.359848484848485
   Second Membership value= Null
   Result = First Membership value.
4. By applying equation 6:
   Likes Ratio = 1, because current Max Likes =0.
5. By applying Mamdani result for Likes Ratio (Equation 9):
   First Membership value= 1
   Second Membership value= Null
   Result = First Membership value.
6. By applying equation 7:
   Account Attitude = 0, because the current count of sentences is 0.
7. By applying Mamdani for Sentence Polarity (Equation 9):
   First Membership value= 0

Second Membership value= Null
Result = First Membership value.

8. COG is applied to get the final fuzzy value (equation 8):
$$Sentence\ Polarit = \frac{0.071969696969697 + 1 + 0}{0.071969696969697 \times 0.799242424242424 + 0.071969696969697\ 1\ \times 1.79924242424242 + \times 0.99999999999999978 + 0 \times 0} = 1.18596236264144$$
9. By applying Mamdani for the COG result (equation 9)
First Membership value= 1
Second Membership value= Null
Result = First Membership value.

Then we start adding one sentence in each round as follow:

First round:

Sentence Text	Processed Text	Polarity	Page	Likes	Fuzzy Result	Creation Time ms
الكونغرس116 يتصدر # قائمة التداول الأمريكية في #تويتر بعد تولّي الديمقراطية نانسي بيلوسي رئاسة مجلس النواب 7	يتصدر قائمة التداول الامريكية تولي الديمقراطية نانسي بيلوسي رئاسة مجلس النواب	0.723167281112344	الجزيرة	25	Very Positive	720000
وزارة الخزانة الأميركية الدين الأميركي زاد أكثر من تريليوني دولار منذ تولي الرئيس #ترامب مهام منصبه في البيت الأبيض 7	وزارة الخزانة الاميركية الدين الاميركي زاد اكثر تريليوني دولار تولي الرئيس مهام منصبه البيت الابيض	1	الجزيرة	79	Very Positive	850000

Table 4-11 Tracking continuous learning - 2 sentences

In this round, the first sentence polarity was changed from 1 to 0.723167281112344

Mathematical explanation:

1. The sentence polarity remains the same.
2. By applying equation 6:
Likes Ratio = 25/79 = 0.3164556962025316.
3. By applying Mamdani result for Likes Ratio (Equation 9):
First Membership value= 0.683544303797468

Second Membership value= -0.316455696202532
Result = Second Membership value.

4. By applying equation 7:
   Account Attitude = 1, because the current count of sentences is 2 and all of them are positive.
5. By applying Mamdani for Sentence Polarity (Equation 9):
   First Membership value= 1
   Second Membership value= Null
   Result = First Membership value.
6. COG is applied to get the final fuzzy value (equation 8):
   $$Sentence\ Polarit = \frac{-0.316455696202532 + 1 + 0.071969696969697}{-0.316455696202532 \times 0.917721518987342 + 0.316455696202532X\ 0.841772151898734 + 1 \times 0.999999999999999 + 0.071969696969697 \times 0.799242424242424}$$
   $$= 0.944633456222469$$
7. By applying Mamdani for the COG result (equation 9)
   First Membership value= 0.723167281112344
   Second Membership value= Null
   Result = First Membership value.

Second Round:

Sentence Text	Processed Text	Polarity	Page	Likes	Fuzzy Result	Creation Time ms
الكونغرس116 يتصدر # قائمة التداول الأمريكية في #تويتر بعد تولّي الديمقراطية نانسي بيلوسي رئاسة مجلس النواب 7	يتصدر قائمة التداول الامريكية تولي الديمقراطية نانسي بيلوسي رئاسة مجلس النواب	0.723167281112344	الجزيرة	25	Very Positive	720000
وزارة الخزانة الأميركية الدين الأميركي زاد أكثر من تريليوني دولار منذ تولي الرئيس #ترامب مهام منصبه في البيت الأبيض 7	وزارة الخزانة الاميركية الدين الاميركي زاد اكثر تريليوني دولار تولي الرئيس مهام منصبه البيت الابيض	1	الجزيرة	79	Very Positive	850000
اعتماد الدول الأوروبية أو ارتباطها بعقود اقتصادية مع دول الخليج يعيقها عن التدخل في الشؤون الخليجية وتدخلها في الصر 8954	اعتماد الدول الاوروبية او ارتباطها بعقود اقتصادية دول الخليج يعيقها التدخل الشؤون الخليجية وتدخلها الصر	1	الجزيرة	59	Very Positive	710000

Table 4-12 Tracking continuous learning - 3 sentences

Nothing was changed in this round for the previously gathered sentences because the sentence did not change, the maximum likes' value was not changed, and the account orientation was not changed because all of the sentences are Very Positive.

We tracked the first sentence classification changes in the chart below:

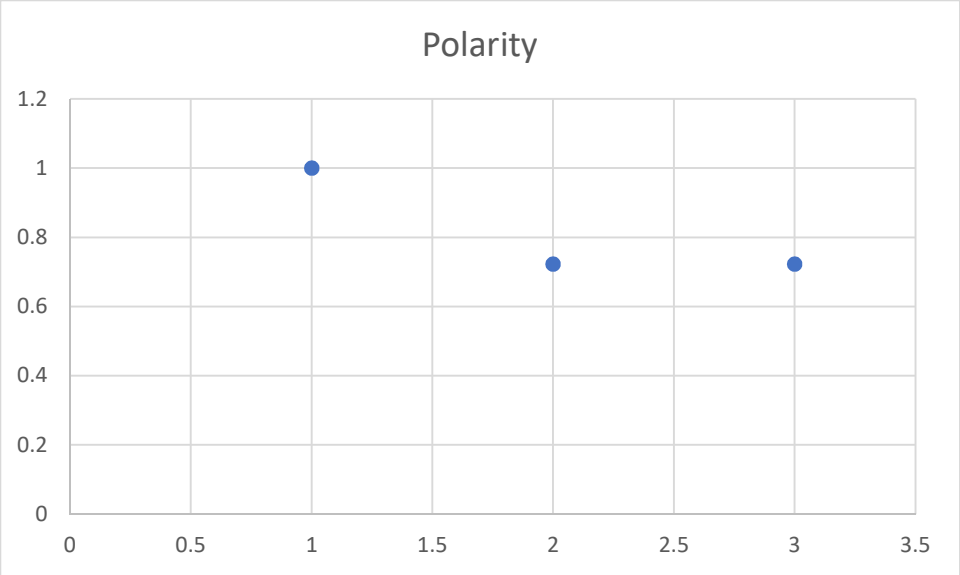

Figure 4-7First sentence classification changes

4.4 Results Discussion

The results presented in section 5.2 show the actual behavior of the algorithm in classifying the obtained data; in one hand, the results show that each obtained sentence is gaining an initial classification that might not reflect its actual classification, but on the other hand, the accuracy of the classification is increasing more and more when more sentences for the same account are gathered and classified.

The second round on any sample changed the classification in that sample, but third and fourth rounds on the same sample will not change the classifications in the sample due to the fact that no new sentences were inserted to the sample so there was no new input that the system will learn from them and change the sample classification based on it.

The system is continuously learning and changing the classification if and only if there are new input data to update the Likes Ratio and the Account Orientation Ratio due to the fact that the third item in the classification formula which is the Sentence Polarity is not changing and it is a constant value and calculated only for one time for each sentence.

The scalability of the system was summarized in scalability table in section 5.4, the memory consumption remains static which is reasonable because the system used 100% since the first round; CPU and time consumption reduced several times due to saving translations results in the database which reduce the need to access Google translate and that saved time and reduced the CPU usage.

The figure below is illustrating the time consumption in each iteration, the figure 22 is showing a reduction in the execution time from the first round to the second round and another reduction from the fourth round to the fifth round.

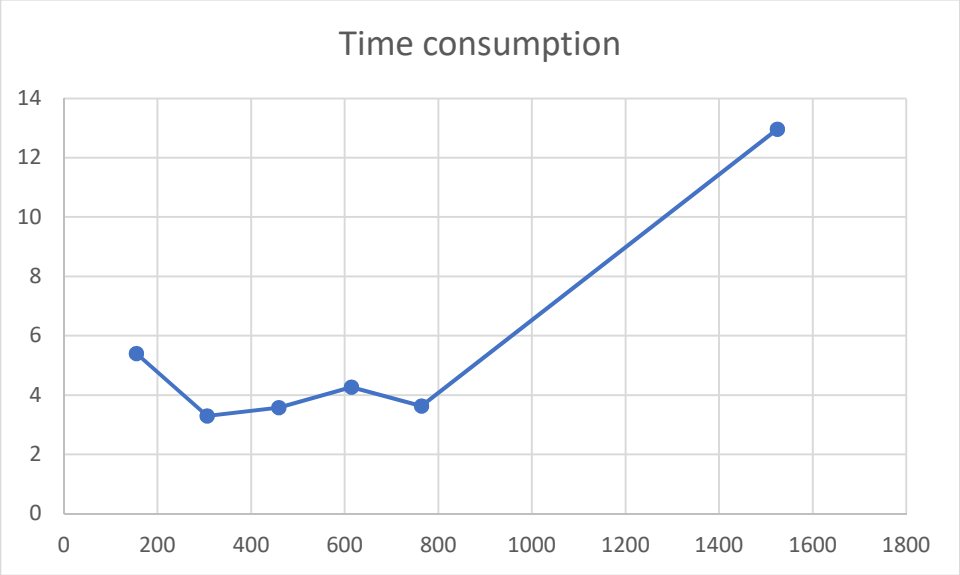

Figure 4-8 Time consumption during scalability testing

The figure below is illustrating the CPU consumption in each iteration, the figure 23 is showing a reduction in the CPU consumption from the first round to the second round and another reduction from the fourth round to the fifth round.

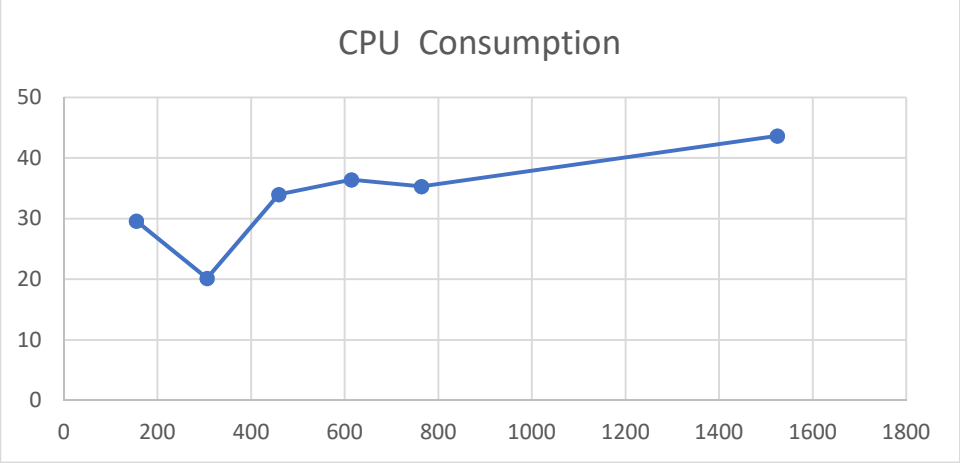

Figure 4-9 CPU consumption during scalability testing

The figure 24 below illustrates the CPU consumption and the time consumption in each iteration, the figure is showing that the reduction in the CPU consumption is synchronized with the reduction in execution time and vice versa.

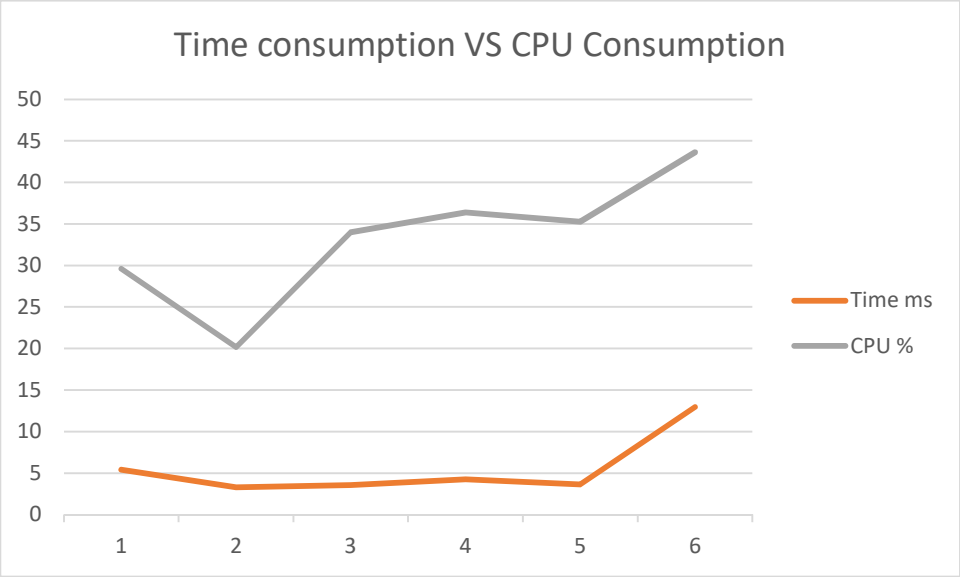

Figure 4-10 Time consumption VS CPU Consumption

4.5 Summary

In this chapter, we show the collected data and data sources for two rounds, then we show the changes in the collected data in each cycle and we prove the continuity of learning in the proposed system.

In addition, we show the system scalability and we end with the results discussion to explain the obtained results.

Chapter 5: Conclusion

5 Conclusion and Future Work

The sentiment analysis essential objective is to classify the polarity of a text if it is positive, negative, or neutral, additionally, it can recognize the emotional state. The complexity in sentiment analysis exists in as data collection, data processing, and the nature of the data.

In this research, we proposed a new sentiment analysis system for classifying the textual content from social media into seven classifications which. A state of art is provided, it discussed the related work in four main fields; dictionary approach, fuzzy logic approach, sentiment analysis on social media, and the lexicon analysis. Sentiment analysis is studied in details; this is in order to provide the most suitable solution that can address the most limitations in the system.

In the proposed system, applying sentiment analysis to Classical Arabic was the main challenges. CAAS succeeded in solving all the well-known sentiment analysis complexity issues; the system succeeded in collecting data from Twitter, which was done using Twitter's API; also, the system succeeded in applying the pre-process cleanup procedures, the sentences were clean and suitable for analysis; the main succeed was in implementing and testing the hybrid core, it consists of dictionary and fuzzy logic.

The proposed system approved that the general attitude has an influence on the sentence sentimental classification, also, it approved that the sentence popularity has an influence on the sentence sentimental classification, those effects were calculated by the system and were used as parameters in the center of gravity formula. Also, we show how gathering new sentences is affecting the previously collected posts polarities and classifications.

While Facebook was one of the targeted data sources, we could not collect data from it during the testing phase according to changes in Facebook privacy conditions. The implemented model for collecting data from Facebook was working successfully during the implementation and code testing phase.

The future work will be enhancing the proposed solution in several aspects, we will design and implement a new model for collecting data from Facebook, also, we will work more on the ontological part to apply the effect of the ontology on the classification of the sentences, and finally we will work on dialects and Arabizzi to have a complete solution for any form of Arabic language.

6 References

[1] E. Cambria, " Affective computing and sentiment analysis," *IEEE Intelligent Systems,* pp. 102-107, 2016.

[2] M. Wen, D. Yang and C. P. Rosé, "Sentiment Analysis in MOOC Discussion Forums: What does it tell us?," in *7th International Conference on Educational Data Mining*, London, 2014.

[3] M. Sadegh, R. Ibrahim and Z. A. Othman, "Opinion Mining and Sentiment Analysis: A Survey," *International Journal of Computers & Technology,* vol. 2, no. 3, June 2012.

[4] M. Dragoni, A. G. B. Tettamanzi and C. . D. C. Pereira, "Propagating and Aggregating Fuzzy Polarities for," *Cognitive Computation,* pp. 186-197, July 2015.

[5] B. Liu, Sentiment analysis and opinion mining, vol. 5, Williston: Synthesis lectures on human language technologies, 2012.

[6] L. S. Larkey, L. Ballesteros and M. E. Connell, "Improving Stemming for Arabic Information Retrieval:," in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, 2002.

[7] A. Rajadesingan, R. Zafarani and H. Liu, "Sarcasm Detection on Twitter: A Behavioral Modeling Approach," in *WSDM '15 Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, Shanghai, China, 2015.

[8] R. Rajnish, "Fuzzy Aspects in Sentiment Analysis and Opinion Mining," *International Journal of Innovative Research in Science, Engineering and Technology,* vol. 7, no. 5, pp. 7750-7755, 2016.

[9] A. Hotho, A. Nurnberger and G. Paaß, "A Brief Survey of Text Mining," *Ldv Forum,* vol. 20, no. 1, pp. 19-62, 13 May 2005.

[10] R. Bhonde, B. Bhagwat, S. Ingulkar and A. Pande, "Sentiment Analysis Based on Dictionary Approach," *International Journal of Emerging Engineering Research and Technology,* vol. 3, no. 4, pp. 51-55, January 2015.

[11] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity," in *42nd annual meeting on Association for Computational Linguistics*, 2004.

[12] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, 2004.

[13] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends® in Information Retrieval, 2008.

[14] S. Baccianella, A. Esuli and F. Sebastiani, "SENTIWORDNET 3.0: An Enhanced Lexical Resource," in *REC 2010, Seventh International Conference on Language Resources and Evaluation*, Valletta, Malta, 2010.

[15] S. Babar and S. Thorat, "Improving text summarization using fuzzy logic & latent semantic analysis," *International Journal of Innovative Research in Advanced Engineering (IJIRAE),* vol. 1, no. 4, pp. l70-177, May 2014.

[16] P. Buitelaar, P. Cimiano and B. Magnini, "Ontology learning from text: An overview.," in *ntology learning from text: Methods, evaluation and applications*, Amesterdam, IOS Press, 2005, pp. 3-12.

[17] W. Liu, M. Bennamoun and W. Wong, "Ontology learning from text: A look back and into the future," *ACM Computing Surveys (CSUR),* vol. 44, no. 4, p. 20, 2012.

[18] M. Abdul-Mageed, S. Kübler and M. Diab, "Samar: A system for subjectivity and sentiment analysis of arabic social media," in *Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis*, Jeju, 2012.

[19] R. . M. Duwairi and I. Qarqaz, "Arabic Sentiment Analysis using Supervised Classification," in *Future Internet of Things and Cloud (FiCloud), 2014 International Conference on*, Barcelona, 2014.

[20] F. Ferr, P. Grifoni, F. Ferri and T. Guzzo, "Approaches, Tools and Applications for Sentiment," *International Journal of Computer Applications,* vol. 125, no. 3, pp. 26-33, September 2015.

[21] P. Mishra, R. Rajnish and P. Kumar, "Sentiment analysis of Twitter data: Case study on digital India," in *Information Technology (InCITe)-The Next Generation IT Summit on the Theme-Internet of Things: Connect your Worlds, International Conference on*, Noida, 2016.

[22] S. A. Salloum, A. Q. AlHamad, M. Al-Emran and K. Shaalan, "A Survey of Arabic Text Mining," in *Intelligent Natural Language Processing: Trends and Applications*, Cham, Springer, 2018, pp. 417-431.

[23] T. Hardeniya and D. A. Borikar, "An Approach To Sentiment Analysis Using Lexicons With," *IOSR Journal of Computer Engineering (IOSR-JCE),* vol. 18, no. 3, pp. 53-57, 2016.

[24] E. Kouloumpis, T. Wilson and J. Moore, "Twitter Sentiment Analysis: The Good the Bad and the OMG!," in *Fifth International AAAI Conference on Weblogs and Social Media*, Barcelona , 2011.

[25] J. Jiao and Y. Zhou, "Sentiment Polarity Analysis based multi-dictionary," *Physics Procedia,* vol. 22, pp. 590-596, January 2011.

[26] A. Kumar and T. M. Sebastian, "Sentiment Analysis on Twitter," *IJCSI International Journal of Computer Science Issues,* vol. 9, no. 4, pp. 372-378, July 2012.

[27] N. Mittal and B. Agarwal, "Sentiment Analysis of Hindi Review based on Negation and Discourse," in *11th Workshop on Asian Language Resources*, Nagoya, 2013.

[28] Y. Sharma , V. Mangat and M. Kaur, "A practical approach to Sentiment Analysis of hindi tweets," in *Next Generation Computing Technologies (NGCT), 2015 1st International Conference on*, Dehradun, 2015.

[29] D. Jhaveri, A. Chaudhari and L. Kurup, "Twitter Sentiment Analysis on E-commerce Websites in India," *International Journal of Computer Applications,* vol. 127, no. 18, pp. 14-18, October 2015.

[30] Z. Liu, N. Yang and S. Cao, "Sentiment-analysis of review text for micro-video," in *Computer and Communications (ICCC), 2016 2nd IEEE International Conference on*, Chengdu, 2016.

[31] D. M. E.-D. M. Hussein, "A survey on sentiment analysis challenges," *Journal of King Saud University – Engineering Sciences,* pp. 330-338, 2018.

[32] S. Zhang, Z. Wei, Y. Wang and T. Liao, "Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary," *Future Generation Computer Systems,* pp. 395-403, 2017.

[33] T. Ahmad and H. Rahmath, "Fuzzy based Sentiment Analysis of Online Product Reviews using Machine Learning Techniques," *International Journal of Computer Applications,* vol. 99, no. 17, pp. 9-16, 2014.

[34] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan and N. A. Smith, "Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments," in *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers*, Stroudsburg, 2011.

[35] R. M. Duwairi , R. Marji, N. Sha'ban and S. Rushaidat , "Sentiment Analysis in Arabic Tweets," in *Information and communication systems (icics), 2014 5th international conference on*, Irbid, 2014.

[36] S. Batabyal, B. Somsukla and B. Somprakash, "Extracting useful information from Facebook posts: A Cognitive-Linguistic Analysis," *International Journal of Humanities and Management Sciences,* vol. 2, pp. 172-178, 2014.

[37] N. A. M. ZAMANI, S. Z. Z. ABIDIN, N. OMAR and M. Z. Z. ABIDEN, "Sentiment Analysis: Determining People's Emotions in Facebook," in *13th International Conference on Applied Computer and Applied Computational Science*, Iwate, 2014.

[38] S. M. Mohammad, M. Salameh and S. Kiritchenko, "Sentiment Lexicons for Arabic Social Media," in *10th Edition of its Language Resources and Evaluation Conference*, Portorož , 2016.

[39] N. Öztürk and S. Ayvaz, "Sentiment Analysis on Twitter: A Text Mining Approach to the Syrian Refugee Crisis," *Telematics and Informatics,* pp. 136-147, 2017.

[40] P. Sojka and A. Horák, "A lexicographer-friendly association score," in *Recent Advances in Slavonic Natural Language Processing*, Brno, 2008.

[41] M. Alrabiah, A. Al-salman and E. Atwell, "A New Distributional Semantic Model for Classical Arabic," in *2nd International Conference on Islamic Applications in Computer Science and Technology (IMAN 2014)*, Amman, 2014.

[42] J. I. Sheeba and K. Vivekanandan, "A fuzzy logic based on sentiment classification," *International Journal of Data Mining & Knowledge Management Process,* vol. 4, no. 4, pp. 27-44, 2014.

[43] F. Moltmann, "Natural language ontology," 2017.

[44] H. Ishkewy, H. Harb and H. Farahat, "Azhary: An Arabic Lexical Ontology," *International Journal of Web & Semantic Technology (IJWesT),* vol. 5, no. 4, pp. 71-82, October 2014.

[45] T. Hardeniya and D. A. Borikar, "An Approach To Sentiment Analysis Using Lexicons With Comparative Analysis of Different Techniques," *IOSR Journal of Computer Engineering (IOSR-JCE),* vol. 18, no. 3, pp. 53-57, 2016.

[46] L. A. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning," *Information sciences,* vol. 8, no. 3, pp. 199-249, 1975.

[47] E. H. Mamdani and P. J. King, "The application of fuzzy control systems to industrial processes," *Automatica,* vol. 13, no. 3, pp. 235-242, 1977.

[48] L. A. Zadeh, "On the validity of Dempster's rule of combination of evidence," Electronics Research Laboratory, College of Engineering, University of California, Berkeley, 1979.

[49] I. Iancu, "A Mamdani type fuzzy logic controller," in *Fuzzy Logic-Controls, Concepts, Theories and Applications*, InTech, 2012, pp. 325-350.

[50] D. M. &. E. N. B. Boyd, "Social network sites: Definition, history, and scholarship," *Journal of computer-mediated Communication,* pp. 210-230, 2007.

[51] J. A. Obar and S. Wildman, "Social media definition and the governance challenge-an introduction to the special issue," *,* 22 July 2015.

[52] H. Kwak, C. Lee, H. Park and S. Moon, "What is Twitter, a Social Network or a News Media?," in *19th international conference on World wide web*, New York, NY, 2010.

[53] M. Thornton, Z. Kalibala, E. Ewemie and H. Alim, "What is Facebook?," [Online]. Available: http://www.academia.edu/download/33463473/Facebook_as_an_Organization.doc.

[54] D. Boyd and K. Crawford, "Six Provocations for Big Data," in *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*, 2011.

[55] M. Beyer, "Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data," 2011. [Online]. Available: https://www.gartner.com/newsroom/id/1731916.

[56] M. Hilbert, "Big Data for Development: A Review of Promises and Challenges," *Development Policy Review, 34,* p. 135–174, 2016.

[57] C. BOJA, A. POCOVNICU and L. BĂTĂGAN, "Distributed Parallel Architecture for "Big Data"," *Informatica Economică,* vol. 16, no. 2, pp. 116-127, 2012.

[58] C. M. Judd and G. H. McClelland, Data Analysis: A Model Comparison Approach, New York: Harcourt Brace Jovanovich, 1989.

[59] C. M. Judd, G. H. McClelland and C. S. Ryan, Data Analysis: A Model Comparison Approach, New York: Routledge, 2008.

[60] C. a. R. S. O'Neil, Doing data science: Straight talk from the frontline, Sebastopol, California: O'Reilly Media, Inc, 2013.

[61] S. E. M. F. U. G. J. H. J. M. S. .. &. W. W. Chakrabarti, "DATA MINING CURRICULUM: A PROPOSAL," Intensive Working Group of ACM SIGKDD Curriculum Committee, 2006.

[62] C. Clifton, "Data mining," 27 September 2017. [Online]. Available: https://www.britannica.com/technology/data-mining.

[63] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth , "From Data Mining to," *AI Magazine,* vol. 17, no. 3, pp. 37-54, 1996.

[64] "Tips on Creating a Social Media Marketing Strategy for Trenchless Business," [Online]. Available: https://trenchlessmarketing.com/2016/11/15/tips-creating-social-media-marketing-strategy-trenchless-business/.

[65] S. Pratt, "5 (and a Half) of the Best Twitter Practices," 31 October 2013. [Online]. Available: https://www.socialtalent.com/blog/recruitment/5-and-a-half-of-the-best-twitter-practices.

[66] R. K. Nielsen and K. C. Schrøder, "The relative importance of social media for accessing, finding, and engaging with news: An eight-country cross-media comparison," *Digital journalism,* pp. 472-489, 2014.

7 Appendix A – Pseudocode

7.1 Data Collection Phase Pseudocode

```
IF Reading Source = Twitter THEN
        Initialize Twitter Access
        Specify The Targeted Page
        Get Tweets From The Page
ELSE
        // If the source is not Twitter it will be Facebook
        Initialize Facebook Access
        Specify The Targeted Page
        Download Encrypted Posts
        Decrypt The Posts
END IF
Insert Sentences To Database
```

7.2 Pre-Processing Phase Pseudocode

```
Get Sentences From Database
Get StopWords From Database
FOR Each Sentence in Sentences
        Remove Punctuations From Sentence
        Tokenize the Sentence
        FOR Each Token in Sentence
                IF Token is StopWord THEN
                        Remove Token From Sentence
                END IF
                IF Token is URL THEN
                        Remove Token From Sentence
                END IF
                IF Token is Number THEN
                        Remove Token From Sentence
                END IF
                IF Token is Non-Arabic Letter THEN
                        Remove Token From Sentence
                END IF
        END FOR
        Insert Processed Sentence In Database
END FOR
```

7.3 Emojis Detection Phase Pseudocode

```
Get Sentences From Database
FOR Each Sentence in Sentences
Emojis Polarity = 0
        Tokenize Sentence
        FOR Each Token in Sentence
                IF Token is Emoji THEN
                        Emojis Polarity += Token Polarity
                END IF
        END FOR
        Store Emoji Polarity
END FOR
```

7.4 Dictionary Phase Pseudocode

```
Get Sentences From Database
FOR Each Sentence in Sentences
        Tokenize the Sentence
        FOR Each Token in Sentence
                Send Token to Google Translate
                Get Translated Token
                IF Translated Token in SentiWordNet THEN
                        Get Token Polarity From SentiWordNet
                        Store Token Polarity
                ELSE
                        Get Token Stem
                        Send Stem to Google Translate
                        Get Translated Stem
                        IF Translated Stem in SentiWordNet THEN
                                Get Stem Polarity From SentiWordNet
                                Store Stem Polarity
                        ELSE
                                Token Polarity = Neutral
                END IF
        END FOR
END FOR
```

7.5 Fuzzy Logic Phase Pseudocode

```
Get Sentences From Database
Get Rules From Database
FOR Each Sentence in Sentences
        //Fuzzification:
        Calculate the Sentence Polarity Fuzzification Value
        Calculate Likes Ratio Fuzzification Value
        Calculate Account Ratio Fuzzification Value
        // Rule Evaluation
        FOR Each Rule in Rules
                IF Rule Applied Then
                        Rule.Apply(Sentence)
                        Calculate COG Value
                        Calculate COG Fuzzification Value
                END IF
        END FOR
        //Defuzzification
        Defuzzify(Store COG Fuzzification)
        Store Defuzzification Value in Database
END FOR
```

الملخص

أصـــبحت شـــبكات التواصـــل الاجتماعي في الســـنوات الأخيرة منجما من الذهب للتحليلات والاستدلالات وبيئة غنية يمكن استغلالها لتطوير المعرفة في مختلف المجالات .
يمكن تصـنيف التحليل الدلالي في وسـائل التواصـل الاجتماعي الى ثلاثة مناهج رئيسـية وهي التحليل الدلالي القائم على المحتوى ، والتحليل الدلالي القائم على المســتخدم ، والتحليل الدلالي القائم على شبكة المستخدم. يركز المنهج الأول على المحتوى النصـي لمنشـورات المستخدمين فقط، بينما يركز المنهج الثاني على المستخدم بحد ذاته من حيث تكرار انماط الاستخدام والميول الخاصـة بالمسـتخدم، بينما يركز المنهج الثالث على شـبكة المسـتخدم والتي تتضـمن الأصـدقاء والمتابعين والحســابات المتابعة من قبل المســتخدم بالإضــافة إلى المحتوى المشــارك من قبل المستخدم. في هذا البحث سوف ينصب اهتمامنا على المنهج الأول فقط.

هنالك عدة خوارزميات اســتخدمت للوصــول إلى أفضــل دقة ممكنة في التحليل الدلالي على مســتوى شــبكات التواصــل الاجتماعي، كان أكثرها دقة هو اســتخدام منهج هجين يمزج بين القواميس والمنطق الضـبابي وهو ما اسـتعملناه في هذه الأطروحة للحصـول على أفضـل النتائج الممكنة.

كخلاصــة للنتائج، فقد قدمنا نموذج أوليا قادرا على حســاب وزن الجمل وتصــنيفها الى ســبع تصــنيفات وهي الإيجابي جدا، الإيجابي، الجيد، الطبيعي، غير الجيد، الســلبي، والســبي جدا، والنموذج المطروح يتمتع بالقدرة على التعلم وتصـحيح نتائجه بناء على المعلومات التي يجمعها و يحللها، وقد قمنا بإثبات صحة حساباته رياضيا.