Master's thesis:

Sentiment Analysis for Lebanese Arabizi Customers' Reviews

Marwan Mohamad Al Omari

Lebanese University

Center of Language Sciences and Communication

Supervised by Dr. Moustafa Al-Hajj

2018-2019

Table of Contents

List of Figures and Tables

Abstract

Because of the huge amount of data that users generate on the web, it is crucial to build a sentiment analysis tool for the Arabizi language system to recognize feelings associated to these data. We proposed two kinds of approaches based on supervised machine learning (ML) using logistic regression (LR) and the other based on lexicon-based using Science of Language and Communication Semantic Analysis System (SLCSAS) tool. Both approaches have been conducted and tested on a corpus of 2635 public and public services' reviews, including hotels, restaurants, shops, governmental institutions (municipalities, universities, and offices, etc.) and other categories, collected from Facebook, Google and Zomato platforms. The total reviews of private service sector are 2501, which overrepresent the sample than the rest 134 reviews of public sector.

At first, data of text reviews have been preprocessed and filtered by 1) removing user's information, 2) transforming texts to lower case, 3) splitting data into 80% training and 20% testing sets, 4) removing reviews with neutral class, and encoding reviews with 0s (negative) and 1s (positive) classes. Then, data feature is considered through BoW and TF*IDF enhanced with word level n-grams dictionary mainly unigrams, bigrams, and trigrams. In SLCSAS, dictionary, grammar rules, and semantic map have been constructed for further implementation. On the other hand, ML models built through the consideration of two phases. The first phase considers the construction of two LR models with default settings set by *scikit-learn* library. The second phase considers using a pipeline to facilitate the hyperparameter tuning of two other LR classifiers. Finally, the results of the five built classifiers are evaluated in terms of precision, recall, f1-score, confusion matrix, and receive operating characteristics curve.

At last, findings show that both LR models trained through BoW and TF*IDF features with default settings remarked similar results, while the hyperparameter tuning of the LR model trained through TF*IDF has surpassed the one with BoW. Therefore, the best nominated classifier is the hyperparameter tuned TF*IDF LR with word level unigram. In addition, SLCSAS classifier has achieved a competitive result in comparison to ML models but with lower coverage on the test data. The ML models so exceed in performance against lexicon-based classifier.

***Keywords**: Sentiment Analysis, Natural Language Processing (NLP), Arabizi NLP, Classifier, Big Data, Machine Learning Classifier, Lexicon-based Classifier, Customer Review.*

Abbreviations and Acronyms

In this section, abbreviated and acronym terms, which used in this paper, are clarified and presented as follows:

Application **P**rogramming **I**nterface (API)

Arabic **N**ame **E**ntity **R**ecognition (ANER)

Area **U**nder the **C**urve (AUC)

Artificial **I**ntelligence (AI)

Automatic **L**anguage **P**rocessing (ALP)

Automatic **L**anguage **P**rocessing **A**dvisory **C**ommittee (ALPAC)

Bag **o**f **W**ords (BoW)

Computational **L**inguistics (CL)

Computer-**m**ediated **C**ommunication (CMC)

Deep **L**earning (DL)

Information **E**xtraction (IE)

Information **R**etrieval (IR)

Logistic **R**egression (LR)

Machine **L**earning (ML)

Modern **S**tandard **A**rabic (MSA)

Natural **L**anguage **P**rocessing (NLP)

Opinion **M**ining (OM)

Part **o**f **S**peech **T**agging (POST)

Receive **O**perating **C**haracteristics (ROC)

Regular **Ex**pressions (REGEX)

Science of **L**anguage and **C**ommunication **S**emantic **A**nalysis **S**ystem (SLCSAS)

Semantic **O**rientation (SO)

Sentiment **A**nalysis (SA)

Term **F**requency **a**nd **I**nverse **D**ocument **F**requency (TF*IDF)

Traitement **A**utomatique du **L**angue (**TAL**)

Transformational **Rules** (T-rules)

World **W**ide **W**eb (WWW)

Personal Motivation

I have been attracted to Natural Language Processing (NLP) expertise area since I was in the third year of English language studies in linguistics branch at the Lebanese University. I was between hundreds of students learning to represent sentences through dependency trees on white boards without the use of technology in structural linguistics course, given in structural linguistics course. Since that, I was fascinated about the automatic language processing (ALP)/traitement automatique du langue (TAL) and started extensively to search for available opportunities in professional research.

Even with the barriers of programming back in the mind of intelligence applications, I became much burning with enthusiastic knowledge. Every day, I learn more and more in the fast-technological world. Much practically, Sentiment Analysis (SA) has caught my attention to further my research on. Because users' opinions in the virtual could take much time and effort to analyze by human hands. Machine doing the analysis with a glimpse of an eye is greatly mind-driven and time-saving. I put my mind on SA because I believe this task enfolds the core of artificial intelligence (AI); and it is highly important in sentiment predication for future happiness and statistician of a nation based on available data in the web. In other words, governments could know and ensure what whole nations would feel and reckon in the nearest and furthest future through deep SA.

Accordingly, I am motivated to analyze texts in theoretical and applicable studies on SA. As a contribution of me with my research team Dr. Moustafa Al-Hajj and Dr. Amani Sabra, we have researched on SA and NLP applications in the Arabic language (Al Omari and Al-Hajj, 2019; Al Omari et al., 2019; Al Omari, Al-Hajj, and Sabra, 2019). Most importantly, we have reviewed lexicon-based, ML, and deep learning (DL) classifiers in wide variety of NLP tasks and applications on SA, sentence categorization, part-of-speech-tagging, language identification, name entity recognition, authorship attribution, word sense disambiguation, and text classification (Al Omari and Al-Hajj, 2019). We have indicated recent challenges in Arabic language processing with further solutions: 1) solid training on NLP approaches of lexicon-based, ML, and DL, 2) accessibility to research materials, 3) increasing the fund to research development.

Furthermore, we proposed LR model trained with term and inverse document frequency (TF*IDF) for the sentiment classification of customer reviews in Arabic language on OCLAR

(Opinion Corpus for Lebanese Arabic Reviews) dataset (Al Omari et al., 2019). Results of the experiment have remarked the unbalance representation of OCLAR dataset; therefore, the classifier's confident was low in sentiment predication. In total, the f-measure is 0.15% on negative class and 0.94% on positive class. In following research, we structured DL architecture of both convolutional neural network (CNNs) and long-term short memory (LSTM) in cope with state-of-the-art literature (Al Omari, Al-Hajj, and Sabra, 2019). The architecture has achieved the state-of-the-art performance on three benchmark datasets (Sub-AHS, ASTD, and OCLAR) out of two (Main-AHS and Ar-Twitter). The model has achieved the following results in accuracy measure on the five datasets: 0.881 on Main-AHS, 0.968 on Sub-AHS, 0.842 on Ar-Twitter, 0.7918 on ASTD, 0.903 on OCLAR.

Scientific Motivation

The research study is practical in the Arabizi language system because there are only few studies dedicated to that system. This research investigates the language components of Arabizi through manual and analyses of services reviews.

First, this thesis works on extracting keywords from the text reviews in way to ease the process of sentiment classification by categorizing significant words into classes. In other words, the words that indicate the classification class of the text under study. The hand-made keywords are the core of lexicon-based analyzer. The SLCSAS classifier is being used to experiment all the maps between keywords in variety of classes. This study remarks significant classification approach and methodology that is well constructed to classify reviews in Arabizi language system. Also, it would be greatly resourced in the general evaluation of services in Lebanon.

Secondly, the research compared the lexical approach to ML model, which has not been addressed in the literature before. The research remarks the first experiment of LR in classifying Arabizi texts. In addition, the scientific significance of this research also relies on the hyperparameter tuning of LR model in goal to reach optimal performance. Therefore, this research leaves many questions and further problems to deal with in upcoming research studies.

Acknowledgments

I wish to say my sincere regards and thanks to Dr. Moustafa Al-Hajj and Dr. Amani Sabra who have been a great research team and supervising me during the master's thesis on SA for Arabic and Arabizi language systems in the Lebanese Context. It is worth to mention the great help, encouragement, guidance, and support I received from both all the time.

Moreover, I wish to express my thanks and gratitude to Coursera's learning platform because it gives me the opportunity to explore the knowledge beyond the physical boundaries. Also, I would like to thank my brothers and my mother for the constant support and encouragement I have received throughout my journey.

I. Introduction

Nowadays, the huge flow of unstructured (unlabeled) data of about forty thousand exabytes that speculated to reach in the early 2020 (Gantz and Reinsel, 2012), with the presence of the World Wide Web (WWW), has attracted a large number of data mining researchers for the aim of extracting vivid knowledge and other useful information for making sense of what the people feel and reckon in the virtual space (Waters, 2010). For such big data analysis, Sentiment Analysis (SA) or opinion mining (OM) is a major concern for opinion analytic and extraction from sequences of texts in forms of reviews, discussions, and blogs (Pang and Lee, 2008). SA is one of multidisciplinary research field that includes NLP, Computational Linguistics (CL), Information Retrieval (IR) or Extraction (IE), ML, DL, and Artificial Intelligence (AI) (Feldman, 2013). Concerning emotion understanding and identification in the depth of Computer-mediated Communication (CMC), SA is most practical and useful to carry on because it fills the gap between machine's understanding and human natural language by giving it the ability to identify and grasp sentimental information through written expressions associated within the big data by classifying and processing language and utterance into one of SA predefined classes, for example, positive, neutral, or negative one (Duwairi et al., 2016).

One of the most popular and used social media application in the Arab world is Facebook[1]. It shows a continuous increase in its users, reaching to about 116 and a half million in the Middle Eastern countries, and specifically 360 thousand in Lebanon solely at the beginning of 2018 ("Middle East Internet Statistics", 2018). Accordingly, users generate continues flow of data in every day's basis that are characterized as growing mountains fueled with opinions: reviews, ratings, recommendations, and other useful information (Wright, 2009), especially on public and private services including food, education, hotel, resort, product, shop, and restaurant, etc. (Agarwal et al., 2015). However, various-shaped challenges associated within the folds of the generated big data while attempting to automatically process such in NLP tasks for the sake of knowledge-making and further decision-making in terms of data size, language dialect, and the

[1] https://www.facebook.com

complexity of linguistics form and nature (phonology, morphology, syntax, semantic and pragmatic) (Elgendy and Elragal, 2014).

With the extensive amount of data available online, the style of language writing could be differentiated from one user to another, based on the used language writing style and the educational background. According to the survey of the Arab Social Media Report (2017), 30% of people in the Middle East review public and private services by writing Arabic characters, and 26% of users use Latin characters, while other 15% use a combination of both Arabic and English texts whenever they would like to express their thoughts. Hence, the mixing of more than one language system and Latin characters paves the way for popularizing new ways of writings on the web in CMC settings (Aboelezz, 2012). One of these styles of writings is Arabizi, which is the target language system for automatic treatment and processing for SA task in this research system because of its high use and popularity in Lebanon.

Arabizi is "a slang term (slang vernacular, popular informal speech) describing a system of writing Arabic using English characters. This term comes from two words "arabi" (Arabic) and "Engliszi" (English). The actual word would be "3rabizi" if represented in its own system," (Yaghan, 2008, p. 39). To confirm the use and the existence of Arabizi language system, the results of American University survey in Cairo (2011), on Arabizi language system use by 70 Arabic users in Facebook, notes that more than 82% of people confirmed using Arabizi, 40% used it most of their times, other 20% use it always, and 17% respondents, who does not used it, said it was out of respect for the Quran and as part of effort to keep away their Arabic identities from the Westernized influence of Arabizi. Besides, Yasmine Dabbous argued, in article *Saving Arabic* ("Saving Arabic" ,2014) declared on the UN Arabic Language Day, Arabizi in the past used as to facilitate communication between young users, although nowadays it is being used in notes-taking at the universities in Lebanon, which recognizes a step ahead in its phasal development and a threat to the native language Arabic. Accordingly, Arabizi remarks a case of new-developed language system that could transfer any message as same as Arabic non-Latin system but at the boundaries of CMC settings.

This research thesis focuses on developing as well as deploying efficient and proficient OM model for automatically processing Arabizi language system in the context of both public and private customer service providers in Lebanon country. Service providers include restaurants,

hotels, shopping centers, governmental institutions, etc... Arabizi corpus of 2635 text reviews, which is essential for the building of the OM model, was gathered through crawling pages of service providers in Facebook, Google[2], and Zomato[3] websites over a period of time from April 4, 2018 to October 30, 2018. The research methodology is experimental, quantitative, and descriptive in nature following both supervised ML and lexicon-based approaches, which were used to build the sentiment classifier, composing additional features that would be beneficial for the training and testing of both. On one hand, ML Features include both Bag of Words (BoW) and Term Frequency and Inverse Document Frequency (TF*IDF) paired with word level n-grams in range of unigrams, bigrams and trigrams. ML approach presented by LR, which fits for binary classification problems as in our case, trained on 80% of data and tested on the 20% in two separate specifications with word level n-grams (unigrams, bigrams and trigrams), the first trained on BoW and the another trained on TF*IDF in default settings set by *scikit-learn* library[4], while the other phase considers training two LR models with having their hyperparameters tuned in attempt to reach the optimal performance. The language analysis is automated in ML through data feature representation of text reviews in BoW and TF*IDF. On the other hand, the rule-based classifier includes semantic map categories of delivery, customer service, price, recommendation & suggestion, administration, service, product, market, ambiance, and overall. Besides, dictionary and grammar parser are harvested from 25% of the training set by deep manual language analysis for the purpose of structuring the SLCSAS classifier. In experiment, it is tested on the same testing set that made in LR model.

1.1 **Problem Statement**

Because of the unavailability of SA tools for automatically processing Arabizi language system, building a one is of highly importance. For this, Arabizi language system would be in a place of recognition in the SA field with the increase of internet users, who currently use it and would use in the future. In addition, it would help companies, institutions, small businesses in

[2] https://www.google.com/maps/@34.0427069,35.6546808,8.7z

[3] https://www.zomato.com/lebanon

[4] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

extracting sentiments of positives and negatives much more efficiently in text reviews written in Arabizi; therefore, they would reflect on enhancing the qualities of their provided services.

1.2 **Purpose of the Study**

The main aim of this research thesis is to give credit to the Arabizi language users' feelings and thoughts in Lebanon territory by extracting sentimental knowledge out of expressed sequences of texts in positive or negative impressions. In addition, it is necessary to highlight the challenges that underpin this language system for the public and researchers most particularly to further their research studies on. Moreover, it is crucial to distinguish Arabizi, particularly in the Lebanese context; therefore, it would be a startup point for other researches to build on. Furthermore, this research experiments the machine capabilities on tasks for sentiment predication and classification in the Lebanese Arabizi. And, this thesis is purposeful to build a dataset that contains reliable Arabizi reviews, which could be used for further researches. Researchers could be working on the expansion of this corpus, too. In general, it is important to classify the outstanding number of Arabizi sentences, which could be of great help for media offices, government centers, research facilities, and start-ups businesses in knowledge-making and future current-based predication tasks.

1.3 **Research Questions**

The research questions of the study are:

1. What would be the best approach that fits the SA task?
2. Which data preprocessing techniques are suitable and available for natural language texts?
3. Which data features most fit the trained ML models?
4. Which LR model would be the best for SA task?

To answer the following research questions, experiments are necessary to conduct using different ML models as well as the lexicon-based classifier aligned with diverse data preprocessing steps and features extraction.

1.4 **Research Hypotheses**

The hypotheses for the study are:

1. ML models could classify Lebanese Arabizi texts in binary classes: positive or negative.

2. Dataset could construct from social media sites on pages of public and private services' providers.
3. Private sector services' providers exceed public (governmental) ones.
4. All experiments would show a very slight difference regarding the achieved results.
5. ML implementation exceeds rule-based classifier's performance.

   To whether validate the hypothesis or not, research experiment and data analytic are crucial to carry on in the further sections of this thesis.

1.5 **Significance of the Study**

The significance of this research is to experiment the effectiveness of applying lexicon-based classifier as well as ML algorithms in the benefits of SA for the Lebanese Arabizi, so to suggest improvements for further research studies. Many steps should be followed to reach the desirable importance from this research, as follows:

I. Investigating most effective and used data preprocessing and features methods for sentiment classification.
II. Deciding which ML techniques and rule-based ones to use for achieving the optimal result.
III. Using available computational resources for the building of SA tools.
IV. Analyzing the performance of applied rule-based classifier as well as ML algorithm with respect to features' selections.
V. Offering suggestions for further research improvement based on the performance of each undertaken ML model and lexicon-based one.
VI. Deploying the best classifier of the undertaken experiments as a real-time service application for Lebanese Arabizi sentiment classification.

Most importantly, the study's importance relies in the richness of information that provided for the public. As a researcher, I do find the subject interesting for further studies from multiple places and positions for the purpose of building an outstanding automatic classifier for Lebanese Arabizi; therefore, researchers may reach to a proper classifier that could handle the sentiment extraction from any given Lebanese Arabizi text. Moreover, companies, shopping markets, and institutions aiming at improving their visibility services might consider integrating and incorporating SA model into their systems.

1.6 **Limitations of the Study**

The research study has faced many obstacles in its folds that restricts it to the area where it has been researched on. First, because the dataset is strictly texts written in Arabizi script of the Lebanese dialect, classification results would lose its generality and validation if it used to predicate other text reviews written in a different dialect. For instance, the research tackles the problematic of creating a tool for Lebanese Arabizi sentiment classification, but it could not achieve good results in predicating sentiment of Arabizi, which is written in Egyptian dialect. Most importantly, the size of the dataset that has been used in this study is small in comparison to other high-credential studies. This goes back to the unavailability of any annotated dataset in Arabizi. And due to the shortness of time and the heavy pressure during the master's study, the collected corpus has a small number of sequences of text reviews.

1.7 **Challenges of the study**

Many challenges we have been faced with during the research process in the field of SA for Lebanese Arabizi text reviews in both public and private services' providers. First challenge, we have been looking for reviews written in Arabic script but while surveying the social media websites mainly Facebook, we have found the number of reviews written in Arabizi script is dominating alongside the English language. Therefore, we have shifted the language choice to Arabizi because of its highlighted representation and importance for further SA task. Secondly, reviews were collected through manual extraction by accessing page by page looking for the ones written in Arabizi script. Otherwise, the use of automatic extractor is much encouraged but did not find a one. In addition, the Zomato application programming interface (API) regularizes the access of reviews per page only to the top 10 reviews, which are quite frustrating. Google map and Facebook APIs are not formally described in how to crawl for reviews in selected areas and services providers' pages. Therefore, we have left to the manual picking choice, which took a large amount of time of about 15588 minutes ($\cong 259.8\ hours$) to collect 2635 text reviews. Last but not least, analyzing 25% of the reviews in the training set for the favor of building up the lexicon-based classifier was the hardest thing opposed to what was expected, tacking about 2 weeks of continues works. The evaluation of the results took a similar time period to make the comparison happens between it and the ML LR classifier. Finally, the small corpus size of 2635 reviews with

the high variety of writings that exist within them made the biggest challenge for training proficient classifiers to reach for convincing and optimal performance.

1.8 **Research contributions**

This research thesis on OM for Arabizi language system contributes to the research literature in wide scale. First of all, the research contributes with the collected corpus in Arabizi language, which embarks a start for bright future works. The corpus is also separated to distinguished training and testing sets that would be a best candidate to test other future-based experiments on. Through the focus on problems, challenges, and suggestions that have been encountered and offered in this master's project, better performance on SA as well as on other tasks such as part-of-speech-tagging is possible. Besides, an automatic processing ML tool[5] would be deployed in real-time service for Arabizi language treatment on SA. According, it would be a beneficial tool for users as well as researchers to give their thoughts on for further improvements. In addition, for our best knowledge it is the first experiment of LR model on OM for Arabizi script in the Lebanese dialect and for Arabizi in general. Finally, this research project conducts a first-kind experiment of hyperparameter tuning on SA for Arabizi reviews.

1.9 **Key Terms**

1.9.1 **Sentiment Analysis**: SA is referred to as subjectivity analysis, opinion mining, or appraisal extraction, with relationship to computer-assisted emotion recognition and expression (Pang and Lee, 2008). SA studies subjective elements, which are "linguistic expressions of private states in context," (Wiebe et al., 2004). These sentiment elements may come from single words, phrases, or sentences, or even whole documents, which regarded as a single sentiment unit (Turney and Littman, 2003; Agrawal et al., 2003).

1.9.2 **Natural Language Processing (NLP)**: NLP is sometimes referred to CL that is the engineering and scientific discipline that guides the understanding of the humans' languages from computational perspective for the building of machine applications that could communicate and dialogue with humans through AI and thinking (Schubert, 2019).

---

[5] https://sa-arabizi-lb.herokuapp.com

1.9.3 **Arabizi NLP**: Arabizi is the informal use of Arabic language in form of Latin characters, represented in written form only (Yaghan, 2008). It is typed most often typed on mobile phones and computer keyboard marking a significance spread via social media website (Basis Technology, 2012). And, NLP analyzes of the Arabizi language system of wide variations in grammar, dialect, context, and spelling to make meaning and understanding based on the assigned task, for example, SA, in which the machine would be apply to spot feelings associated with Arabizi natural language texts (Duwairi et al., 2016).

1.9.4 **Classifier**: Classifier is "to classically organize things by classes, by categories, to classify according to a classification," (Larousse, n.d.). Explicably, classifier is "A mapping from unlabeled instances to (discrete) classes. Classifiers have a form (e.g., decision tree) plus an interpretation procedure (including how to handle unknowns, etc.). Some classifiers also provide probability estimates (scores), which can be thresholded to yield a discrete class decision thereby considering a utility function," (Kohavi and Provost, 1998). In SA task, classification is based on binary or multi-class predication.

1.9.5 **Big Data**: "Big data is a term applied to data sets whose size or type is beyond the ability of traditional relational databases to capture, manage, and process the data with low-latency. And it has one or more of the following characteristics – high volume, high velocity, or high variety. Big data comes from sensors, devices, video/audio, networks, log files, transactional applications, web, and social media - much of it generated in real time and in a very large scale," ("Big Data Analytics", n.d.).

1.9.6 **Machine Learning Classifier**: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E," (Mitchell, 1997, p. 2).

1.9.7 **Lexicon-based Classifier**: it is an AI application for manipulating and storing data for knowledge extraction based on human-crafted rule sets that indulge human's analysis and inference.

1.9.8 **Customer Review**: "A consumer's opinion and/or experience of a product, service or business. Reviews can be found on specialist websites and on the websites of many retailers, retail platforms, booking agents, and trusted trader schemes (schemes helping consumers to select a

trader)," (Valant, 2015). More specifically, A review is a "critical evaluation of a text, event, object, or phenomenon. Reviews can consider books, articles, entire genres or fields of literature, architecture, art, fashion, restaurants, policies, exhibitions, performances, and many other forms […]. While they vary in tone, subject, and style, they share some common features: First, a review gives the reader a concise summary of the content […]. Second, and more importantly, a review offers a critical assessment of the content […]. Finally, a review often suggests whether or not the audience would appreciate it," ("Book Reviews", n.d.).

1.10 **Research Outline**

The following sections of the thesis are organized as follows:

Section II describes the main keywords in literature review following a forced order on NLP, big data, SA and its applicable approaches of rule-based, ML, and hybrid ones; Arabizi language system and its relation to the Lebanese culture. Finally, it presents the most relevant works of SA for the Arabic language and Arabizi in particular.

Section III highlights the Arabizi corpus that used for sentiment classification and the challenges that we have found while analyzing text reviews. Also, data preprocessing is explained besides the selected features. Moreover, a theoretical part of used classifiers presented in much detail.

Section IV details the used corpus and the process of data filtering and preprocessing techniques that have been conducted on the dataset; then the feature extraction; therefore, feeding the features into the ML model to create the LR classifier. Secondly, an information is given about the used lexicon-based classifier, which depends on handcrafted grammar, dictionary and semantic map.

Section V presents the experiments that conducted to evaluate the approach taken in this research.

Section VI concludes the ideas of the whole research and give insight about future works.

II. Literature Review

This research section highlights the main research keywords by emphasis on the literature background for each in the latest research field advances. The research keywords are discussed in order through drawing relationship between them, including NLP, big data, SA, approaches to SA, Arabizi language system, and the Lebanese language dialect.

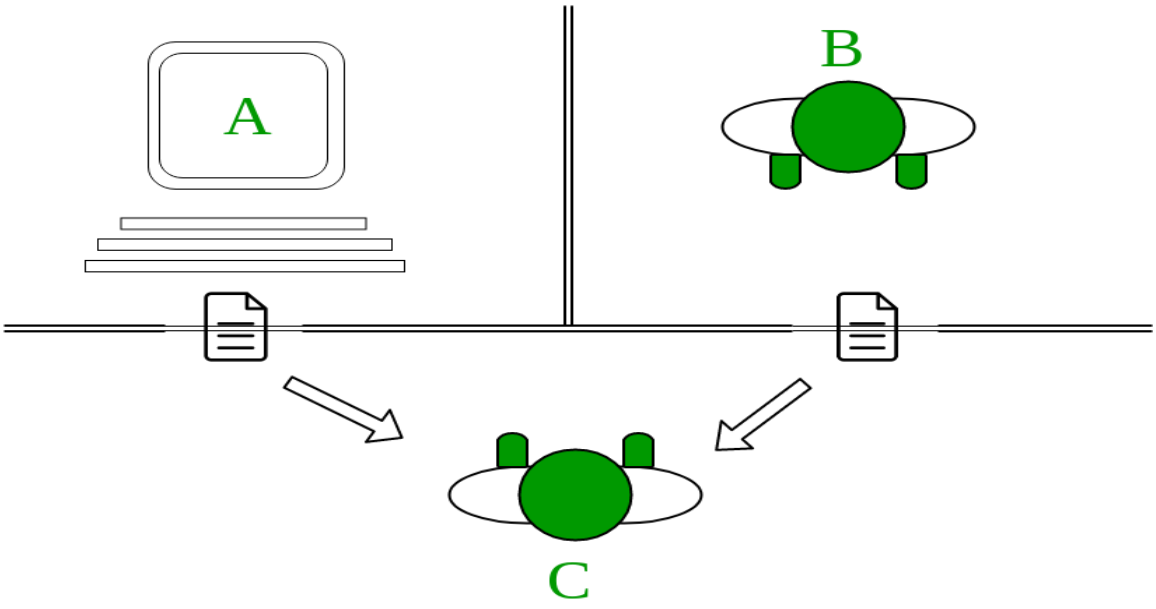
2.1 **Natural Language Processing (NLP)**

NLP focuses on the interactions and communications between machine and human's natural languages by deep process and analysis of natural language data. This field goes way back to the 1950s in which Turing published his famous article titled *Computing Machinery and Intelligence*, which is also known for turning test that is to have a user C chatting in text only through a computer keyboard and screen with other two separate isolated partners, one is a machine A and the other is a human B, as it is presented in Figure (1). Therefore, for the test to be succeeded, the user must be unable to recognize whether he/she is talking to a human or a machine (Saygin, Cicekli, and Akman, 2000). This test challenges the AI breakthrough in robotics, cloud computing systems, and all advances to create a role model of humanized robots that have similar intelligences as humans do. Accordingly, AI does come to alive whenever a single machine A could compete against a human C, as Turning claimed in his research:

> I believe that in about fifty years' time it will be possible, to programme computers, with a storage capacity of about 109, to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning. (Turing, 1950, p.8)

Recently, a program chatbot software, as a talkative 13-year-old Ukrainian lad, successfully passed the imitation game in a competition in the 60th anniversary of Turning's death with a pass mark of 33 that it was a real boy (Williams, 2014).

Figure 1- *Turning Test in which person C chats through text only with another person B and machine A, adopted from Bansal (2018, p.3)*



Later in 1950s, Georgetown experiments on machine translation from Russian to English in which he translated sixty sentences (Hutchins, 2004). It was predicated that within the upcoming five years, machine translation would be at its peak, but the ALPAC (Automatic Language Processing Advisory Committee) in 1966 came up to the spot marking a falling attempt of research on the field (Hutchins and Hays, 2015). Not until 1980s, the field of NLP was an abandonment area of research because of the used approaches based on simple languages' analyses and transformations (Noam, 1965). Language transformational rules (T-rules) are the rules that are applied to change the syntactic structure (deep structure) of sequence of symbols into a new whole syntactic structure that is called a surface structure. These rules are in four forms: deletion, insertion, permutation and substitution.

In 1980s, a transformation period from systems based on hand-written rules to statistical algorithms for NLP introduced such as decision-tree (D-Tree) algorithm and the hidden Markov model (HMM). These models were notably successful in NLP research field as they achieve proficiency results on a large corpus. However, at that time, the major limitation for statistical algorithms were the need for more and more annotated data for computing to the highest performance (Johnson, 2009).

Recent advances in NLP research field mark four main areas of learning techniques: unsupervised, semi-supervised, supervised, and reinforcement learning algorithms that will be elaborated in 2.3.2 **Machine Learning Approach** below. In the 2010s, DL has changed the course of NLP research in the promising state-of-the-art results that could achieve in most of the tasks, including SA (Heikal, Torki, and El-Makky, 2018).

2.2 **Big Data and Sentiment Analysis (SA)**

Crucially, big data mark a big deal for organizations, institutions, and companies in every tiny aspect, ranging from customer information to employees hiring, to product advertising, etc. It shapes the base blocks to build a successful and prosperous business, otherwise business company without big data would be disarmed from extracting any valuable knowledge, performing a strategic decision, as well as knowing market forces of supply and demand (Elegendy and Elragal, 2014). Therefore, organizations must organize big data into a storage for later retrieval. Such methods of storing structured data include database, data mart, and data warehouse by the means of using tools to extract data from external sources, to transform the data to fit the operational needs, and finally to load it into database or data warehouse. Thus, the data made available for data mining and analytical queries (Bakshi, 2012).

On the other hand, data mining works on finding correlations and patterns similarities between the stored big data "databases". It also known as "knowledge discovery in data" (Bastien, 2018). SA is one of the flourishing data mining applications. It focuses on understanding emotions from subjective texts through patterns correlation. It spots opinions and attitudes to a certain topic, subject, and service for the purpose of extracting sentiment knowledge. SA takes NLP as its core for the building of dictionaries that contain sentiment indicators to draw relationship and connection among words in data (Mouthami, Devi, and Bhaskaran, 2014).

SA has multiple analytic levels: document level (Farra et al., 2010), sentence level (Farra et al., 2010; Shoukry and Rafea, 2012a), aspect level (or feature level) (AL-Smadi et al., 2018), word level and character level (Abdualla et al., 2013).

In document level, sentiment is being predicated for the whole document based on the subjectivity expressed within. This method is beneficial to give an overview of subjectivity in each document, but it fails to give a specified knowledge of subjectivity concerning a specific category

as food experience, for example. In this level, sentiment is binary classification problem of positive or negative. Also, it is regarded usually as regression task in which sentiment is 5-stars classification. However, sentiment classification digs deeper into the document when it deals with each sentence entry as specific case in which sentiment is being predicated for each subjectivity or objectivity in a sentence in multiple classes, including positive, negative, or neutral sentiments.

Next, sentiment classification based on word level and character level approach ranks each word in a sentence with a score of positive, negative or neutral class, which is later summed up to obtain the overall sentiment polarity of that sentence. In other words, SA on word level, it deals with each word in an entry as subject for sentimental classification. Overall, the sum of polarities of all words in an entry gives the final sentiment predication.

On the other hand, sentiment with aspect or feature level analyzes a given text by looking into a specific category or feature that text contains. In this approach, sentiment predication identified based on already indicated features. For example, a feature could be the "food experience" in a restaurant. Thus, a review of "Easily the best food in Beirut" would remark a positive sentiment because of "best" indictor for the food experience. This level remarks a good result, but the classifier is only confident whenever it met already-known features, and therefore if the classifier encountered a new text that has unrecognized feature, it would lose its confident. Accordingly, the sentiment classifier would look only on the already trained features unpaying attention to other untrained features within a review. Moreover, this level has two types (Pang, Lee, and Vaithyanathan, 2002):

First, explicit aspect is easily to be identified in a sentence in a form of a phrase. For example, "The food quality is great", here "food quality" is the extracted aspect, which is clearly obvious for annotator. On the other hand, implicit aspect is hard to identify, which requires much reasoning of a sentence. For instance, "The game is not working probably!", here "not working probably" is not the desired aspect but it is the unfunctionally.

2.3 **Approaches to SA**

For the SA task, Various approaches employed including lexicon-based, ML and DL, and lastly hybrid approaches. On one hand, lexicon-based approach makes use of dictionary in the process of either term's inclusive or exclusive from the input text, while ML and DL makes uses

of large dataset for model's training and testing, which then could be used for sentiment predication of a given text. In addition, the hybrid approach takes the output of lexicon-based approach as input for ML, creating a mixture model. The DL, on the other hand, makes a significance hybrid architecture of employing many ML as well as DL models for the purpose of Deep NLP.

2.3.1 **Lexicon-Based Approach**

Many applications nowadays traced to Semantic Orientation (SO), which measures the expressed subjectivity in a document d by either one magnetic side of positive on one side or negative on the other side. This SO approach is considered an unsupervised technique to build the sentiment lexicon by assigning each term t a corresponding class inferred from its semantic intensity in given d (Shoukry and Rafea, 2012b). Thus, terms' classes computed together to formulate the sentiment of that given t (Morsy and Rafea, 2012). This approach sometimes called rule-based approach, which referrers to a handwritten linguistic rule using Regular Expressions (REGEX). However, such system is time consuming; and it requires a long time to adjust (Petasis et al., 2001).

There are two ways to construct sentiment gazetteer or dictionary. First, a one could extract terms manually from a sentence, and annotate them with their corresponding SO. Therefore, dictionary is built automatically by using the terms expansion method where a few seed keywords used to attract other related terms to formulate it. This idea based on the assumption that some terms associate with others that occur with, having a close SO or sentiment polarity (+1 for positive, -1 for negative, or 0 for neutral). According to El-Beltagy and Ai (2013), they provided a clear example illustrating this methodology. It starts by looking on the conjugation patterns of occurred terms in a sentence. For instance, the phrase "محترم و مؤدب" / "polite and respectable", would be regarded as two terms sharing the "same" polarity to certain extent, giving a fact that they are connected grammatically by "و" / "and". As for each expansion, the program user filters the undesired terms manually from the dictionary. These steps are repeated several times until the dictionary reaches its words' counts goal.

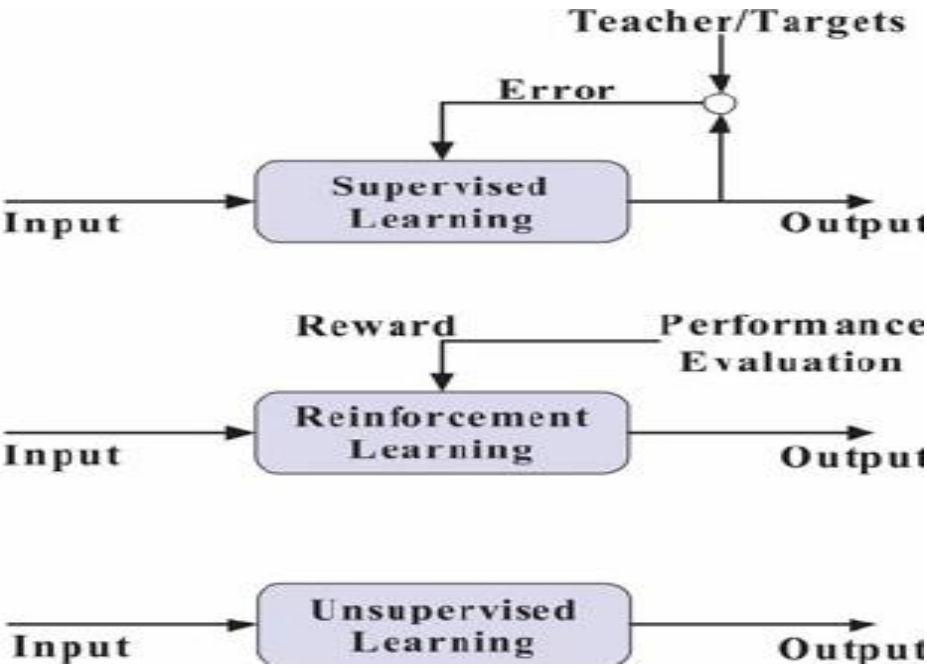
2.3.2 **Machine Learning Approach**

The term ML was first coined by Arthur Samuel in 1959. ML uses statistic as its core so that can learn and construct knowledge from input data to predicate an output (Kohavi and Provost,

1998). ML algorithms undertake data-driven decision and predication on input data (Bishop, 2006). ML employed in a wide range of computing NLP tasks and applications include SA, information extraction (IE), email filtering, summarization, text clustering, question-answering chatbot, and many others (Khurana et al., 2017). In implementation, ML approach takes advantage of the wide range of algorithms and techniques for learning decisions from data for later decision-making. Examples of ML algorithms include decision-tree, LR, and support vector machine, etcetera (Al Omari et al., 2019).

ML algorithms are implemented using four main techniques, which are supervised, unsupervised, semi- supervised, and reinforcement learning, as they elaborated below.

***Figure 2**- ML techniques, adopted from Wang, Chaovalitwongse, and Babuska (2012, p.2)*



2.3.2.1 ***Supervised Machine Learning Technique***

"Techniques used to learn the relationship between independent attributes and a designated dependent attribute (the label)," (Kohavi and Provost, 1998). In this technique, ML algorithms learn decisions and predications from the input (x) and their desired output (y) as to formulate a rule that maps inputs to outputs, as follows:

$$y = f(x) \quad (1)$$

As in later encounter of new input, the algorithm would use what has been learned to predicate its class. This technique is a resemblance of a teacher supervising the learning process of a student, where the algorithm predicates the classes on the training data while it is being corrected by the former teacher (Brownlee, 2016).

2.3.2.2 ***Unsupervised Learning Technique***

"Learning techniques that group instances without a pre-specified dependent attribute," (Kohavi and Provost, 1998). In this technique, ML algorithms learn the pattern similarities among input (x) because there is no desirable output (y) is available. This technique is a resemblance of independent learning where there are neither a teacher nor supervisor to correct the student's work (model's predications). Such applications used to discover the relationships that group data together (Brownlee, 2016).

2.3.2.3 ***Semi-Supervised Learning Technique***

This technique integrates supervised and unsupervised ML Techniques for the tasks where data has only some of unlabeled data (y). It is a practical technique, which saves time and money, makes best predications on unlabeled data as it is useful to use all trained data as an invest on the predication of new untrained data (Brownlee, 2016).

2.3.2.4 ***Reinforcement Learning Technique***

To imitate how human beings learn from their committed errors and mistakes, reinforcement learning allows the machine to learn from committed errors in the predication process to receive a reward in the next time similar encounters (Wang, Chaovalitwongse, and Babuska, 2012). This technique could empower the machine to be an independent learner as if it is a human learning from his/her past experiences.

DL, as a branch of ML, has overcome the baseline ML algorithms with sophisticated architectures and algorithms. A basic architecture of DL consists of an input layer, an output layer, and hidden layer in-between. However, the basic architecture is not referred to as enough deep to be a DL architecture. A deep architecture may consist of more than 3 hidden layers with hundreds of processing units (neurons). DL architectures include, for example, recurrent neural networks

(RNNs), long short-term memory (LSTM), and convolutional neural networks (CNNs), etcetera (LeCun, Bengio, and Hinton, 2015).

2.3.3 **Hybrid Approach**

This Hybrid system blends both previous discussed approaches lexicon-based and ML into one simple mixture to compute in performance score (Petasis et al., 2001). This system starts with performing a lexicon-based approach for the purpose of predicating SO from each sentence in the dataset. Accordingly, the results would be the input of the ML approach, which takes the advantage to learn decision-taking from them. According to Oudah and Shaalan (2012), they worked on Arabic Name Entity Recognition (ANER) using this kind of approach, which has outperformed the state-of-the-art ANER systems with 94.4% f-measure for person named entities, 90.1% f-measure for location named entities, and 88.2% f-measure for organization named entities.

2.4 **Arabizi and the Lebanese Dialect**

For a long period of time before the recognition of Arabic system on the web by means of encoding characters (UTF-8), Arabic users in the internet were forced to create means for communication. Thus, Arabic started to be written in form of Latin characters in what is so known today as Arabizi (Yaghan, 2008), "3arabizi" (Bianchi, 2012), or Romanized Arabic (Al-Khatib and Sabbah, 2008). Those Arabic language users whose native script used non-Latin characters made a language script shift (Crystal, 2001). To mark the correspondence difference between the Arabic encoding characters or letters and Arabizi's, the following Table (1) should be regarded:

***Table 1**- Correspondence Difference Between Arabic and Arabizi Encoding Characters. Adopted from Cotterell et al. (2014, p.2) based on Yaghan (2008)*

Arabic	Arabizi	Arabic	Arabizi
أ	a	ب	b, p
ت	t	ث	th, s
ج	j, g	ح	7, h
خ	7', 5	د	d

ذ	th, z	ر	r
ز	z	س	s, c
ش	sh, ch	ص	9
ض	9', d	ط	t
ظ	th	ع	3
غ	gh, 3'	ف	f
ق	8, 2, k, q	ك	k
ل	l	م	m
ن	n	ه	h
و	w, o, ou	ي	y, i, e

However, the use of letters differs from one person to another based on the background, cultural, and educational preferences (Yaghan, 2008). For instance, ru5am/ ru‟7am/ rukham (marble) shows the use of three different letters to write marble from Arabic to Arabizi. Furthermore, 8arib/ 2arib/ karib (boat). These representations remark different formations of boat in the use of Arabizi letters "8", "2", and "k". Nevertheless, these correspondences highlighted in Table (1) may have similarities according the country region. The following Table (2) shows a demonstration of Arabic and Arabizi corresponding similarities in relation to regional dialects, as it was noted by Tobaili (2015):

Table 2*- Six Similar Arabic and Arabizi Characters in Relation to Five Regional Dialects, adopted from Tobaili (2015, p.3)*

Arabic	**Arabizi**	**Dialect**
جميلة جدا	Jameela Jiddan	MSA
حلوة كتير	7elwe Ktir	Lebanese

حلوة مرة	7ilwa Marra	Saudi
حلوة وايد	7ilwa Wayed	Emirati
حلوة أوي	7ilwa 2awi	Egyptian
جميل برشا	Jmeel Barcha	Tunisian

As it was said, the letter "ح" has a similar corresponding "7" in Arabizi for four dialects. The other two dialects have another word synonym for "حلوة" which is" جميل". It is not the case; however, similarities maybe noted among all dialects, for synonyms exist in nearly most of the Arabic dialect systems.

In addition, Arabizi phenomenon referred to as "slang" (Yaghan, 2008) and "arithmograhemes" in CMC (Bianchi, 2012). According to Széll (2011), "the Arabic we write is not the same as the Arabic we speak," (p.103) and Aboelezz (2012) stated that Arabic is Latinized in the CMC context but still has limited use in formal setting and registers, but in not the case for all Arabic world regions. Clearly, this mixture of Arabizi on one hand and Lebanese nationality on the other hand shows a strong bond (Versteegh, 1997). This relationship could be traced back to Munira Khayyat article in the Daily Star:

> [...] The present extent that English (and French) are being used at the expense of the mother tongue is unprecedented […] nobody in Lebanon seems worried about the invasion of Arab-English (Khayyat, 1999, para. 3 & 9).

That speech validates the relationship between Arabizi and the Lebanese nation, however, it has been a while since then.

2.5 **Sentiment Analysis and Lebanese Arabizi**

Many researches have been tackling sentiment classification for Arabizi in many different dialects. First thing first, Abed AL-Aziz, Gheith, and Ahmed (2011) in a research under title *Toward Building Arabizi Sentiment Lexicon based on Orthographic Variants Identification*, they tackled the problem of sentiment classification by taking benefits of various orthographic representations of Arabizi written in the Egyptian dialect. SoundEX (Bhatti et al., 2014) adopted

from English research to present Specified Arabizi letters by codes. Table (3) shows the corresponding category codes selected for the Arabic letters to Arabizi letters, as follows:

***Table 3**- Arabizi Code Categories, adopted from Abed AL-Aziz, et al. (2011)*

Category code	Arabic	Arabizi Code
0	ا، إ، آ، ح، خ، هـ، ع، غ، ش، و، ي	a, e, a, [7, h], [5, kh, 7'], h, 3, [3', gh], [4, sh, ch], [w, o, u], [y, i, e]
1	ف، ب	[f, v], [b, p]
2	ك، ق، ظ، ز، ص، س، ج، ث، ذ	k, [k, q, 8, 2], [z, 6'], z, [s, 9], [s, c, x], [g, j], th, z
3	ط، ض، د، ت	[6, t], [9', d], d, t
4	ل	l
5	م، ن	m, n
6	ر	r

This code generation helps to standardize the various orthographic representations of a single word in Arabizi. The following Table (4) remarks the unified representation of the word حلو/beautiful in Arabizi:

***Table 4**- Generated Arabizi SoundEX Code for Different Orthographic Forms, adopted from Abed AL-Aziz, et al. (2011)*

Arabizi Word	Generated Arabizi SoundEX Code
gamel	0504
jamil	0504
gameel	0504

jamiiiil	0504
gmeeel	0504
7elow	040
7elw	040
helow	040

In order to evaluate SoundEX code generation, Arabizi sentiment lexicon of 588 colloquial words were annotated as 170 positive words (29.2%) and 412 negative words (70.8%) and written in five different orthographic representation by five native Arabic language users. The proposed methodology achieved 88.67% correct classification and the remain percentage corresponds to misclassification.

Another research by Taha Tobaili (2015) in a thesis titled *Sentiment Analysis for Arabizi in Social Media*, he conducted a lexicon-based approach to classify people's writings in Arabizi on different topics, collected from both Facebook and WhatsApp platforms in the Lebanese coordinates. The used dataset contains 156,040 WhatsApp lines (separated to 14343 negative, 48254 positives, and 93443 neutral classes) and 108,040 Facebook comments (separated to 11494 negative, 45566 positives, and 40980 neutral classes). After the dataset is filtered from senders' names, name tags, links, images, videos, audios, Arabic texts, and emoticons, it feeds into the algorithm which is based on dictionary entity matching. The dictionary is built from a published word list by Hu and Liu (2004). It contains 2000 positive words and 4800 negative words in English language. These words translated into Arabizi in Google Translator[6]. Sentiment classification conducted on word-level polarity, giving each word in a sentence a polarity whether positive, negative or neutral. Overall, sentence classification accuracy achieved 82% for WhatsApp dataset and 77% for Facebook dataset.

[6] https://translate.google.com

Moreover, another research titled *Sentiment Analysis for Arabizi Text* (Duwairi et al., 2016) worked on Arabizi classification into three classes: positive, negative, or neutral classes. Arabizi dataset used for SA that consists of 3206 tweets, labelled as follows: 1803 positive, 831 negative, and 572 neutral sentiments. The dataset has gone through a process of tokenization, emoticons replacement, stop words removal, binary weight, and finally Arabizi letters conversion to Arabic letters. They contributed to the research literature with an achievement of Arabizi letters converter to Arabic language. Every Arabizi word converted into an Arabic word by mapping Arabizi letters to their corresponding in Arabic letters, which was specified in Table (5). The supervised ML technique used both NB and SVM classifiers. Results show that SVM outperforms NB classifier with data filtering by a macro-precision of 0.555 and macro-recall of 0.587; it achieved without filtering 0.549 and 0.584 respectively. On the other hand, NB achieved with filtering a macro-precision of 0.505 and macro-recall of 0.555; it achieved without filtering 0.504 and 0.537 respectively.

Table 5*- Arabizi letters mapping to corresponding Arabic letters, adopted from Duwairi et al. (2016, pp. 129)*

Character in Arabic Language	**Corresponding Character in Arabizi**	**Character in Arabic Language**	**Corresponding Character in Arabizi**
ا	a	ط	T
ب	B	ظ	6’
ت	T	ع	3
ث	t’, th or 4	غ	3’
ج	j or g	ف	F
ح	7	ق	8
خ	5 or 7’	ك	K
د	D	ل	L
ذ	d’	م	M

ر	R	ن	N
ز	Z	ه	H
س	S	و	w or o
ش	$ or sh	ي	e or i
ص	9	ئ	2
ض	9'	ء	2

III. Research Methodology

The research methodology, which is used in this research to approach the sentiment classification for Arabizi reviews written in the Lebanese dialect, uses the quantitative, descriptive and the experimental research methods. The content of this section discusses research design, sample, data preprocessing, data features extraction, and classifiers tools.

3.1 **Research Design**

The research study follows a mixed of quantitative, descriptive, and experimental research designs. First, the quantitative research is statistical in nature collecting numerical data to identify and explicate the effects of using and experimenting diverse ML and rule-based techniques with diverse data preprocessing and features on the overall performance in sentiment classification task. In addition, it is descriptive research used to describe the characteristics of the corpus and the challenges underpinning such SA study. In addition, it is carried for categorizing the text reviews in different elements and classes in SLCSAS classifier. The major variables measured for this study are sentiment ratings, dataset's training and splitting percentages, data preprocessing steps, training data features, and rule-based as well as ML algorithms, and the final evaluation based on the achieved classification results.

3.2 **Research Sample**

This section will present what are the measures and methods that were used to collect the necessary dataset for the study and what are the preprocessing stages that the data go through.

Finally, the procedure of the corpus will be specified as to be readily in service for rule-based classifier and the ML models' trainings.

The target dataset for this research study collected randomly firstly from Facebook on public and private services' reviews inside Lebanon as a whole, including the eight Governorates and districts of Lebanon: Aakkar, Baalbek-Hermel, Beirut, Beqaa, Mount Lebanon, Nabatiyeh, North Lebanon, and South Lebanon ("Territorial administration of Lebanon", 2003). Specific keywords used to spot services providers in Facebook by looking in search bar, using the following examples:

- Lebanon
- Restaurants in Lebanon
- Places to go in Lebanon
- Hotels in Lebanon
- Cafes in Lebanon
- Lebanese Government Organization

Moreover, Zomato (https://www.zomato.com/lebanon) used also to collect reviews from a large catalog of restaurants manually. And, Google Map (https://maps.google.com) as it can be accessed directly by (https://www.google.com/maps/@34.0427069,35.6546808,8.7z) is used to spot reviews about service providers across Lebanon, but the number of reviews were unsatisfactory and small.

***Figure 3**- The Range of Corpus Collection Marked Inside the Dark Line*

Accordingly, Arabizi reviews written in the Lebanese dialect collected from Facebook, Zomato, and Google over a period from April 4, 2018 to October 30, 2018. The corpus has reached to 2635 reviews, generated in Excel CSV (Comma Separated Values) UTF-8 format. A sample from the corpus looks as follow:

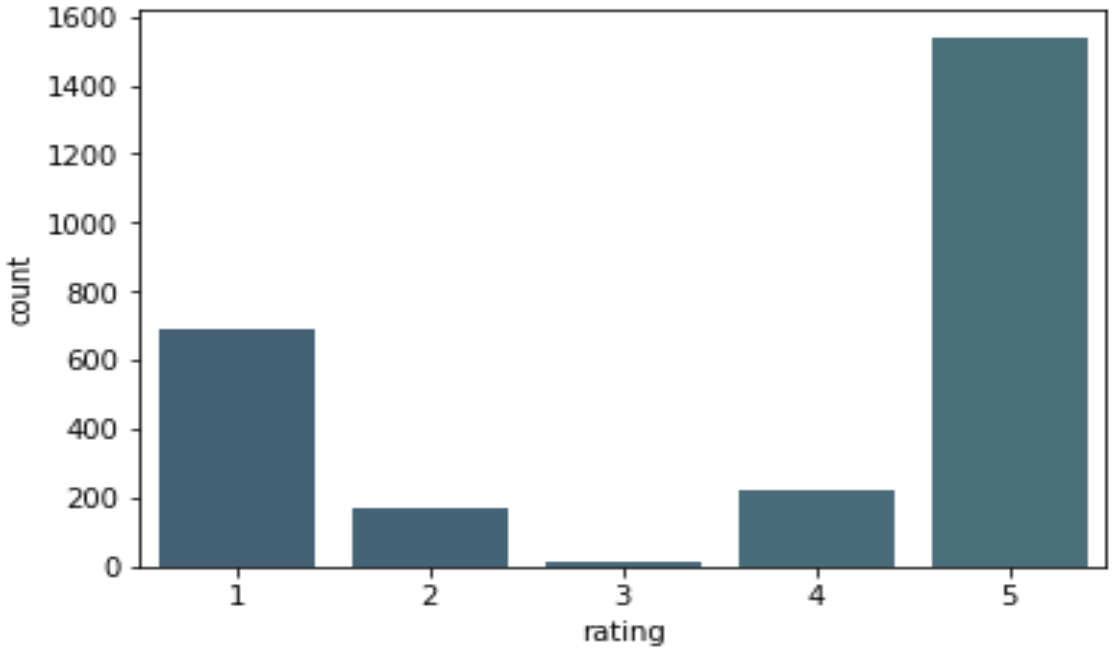
***Table 6**- A sample of the Lebanese Arabizi Corpus*

pagename	servicesector	review	review (in English)	rating
Al Saha Restaurant	private	La3de 7ilwe bs lakel mano tayyib	Beautiful ambiance but the food is not delicious	2
Al Jawad Restaurant	private	sayer abadan ma tayib	it is no more delicious	1
Shadi Najjar Store	private	Walla mnih	Wallah good	5

Hayat Doner Alturki	private	Ktir tayeb	Very delicious	4
Ahwak	private	7aramiye	Thieves	1
Al Naqoura	public	Raw3a ajmal mant2a be lbnen alla yhmeha w dalla 3roset al janoub	Wonderful the most beautiful area in Lebanon, Allah may protect it and stays the south's bride	5
Lebanese University	public	Unprofessional team ! ma fi mozakara 2ella wfiya ghalat ! da7akto el 3alam 3laykon !	Unprofessional team! There is no quiz without erros! You let people to laugh at you!	1
Lebanese University - Faculty of Science	public	A7la uni jad bl ro8em mn kl do8outat li fiya bas a7la jem3a bl kon :) :)	The most beautiful university despite all the pressures, but the most beautiful university in the world :) :)	5

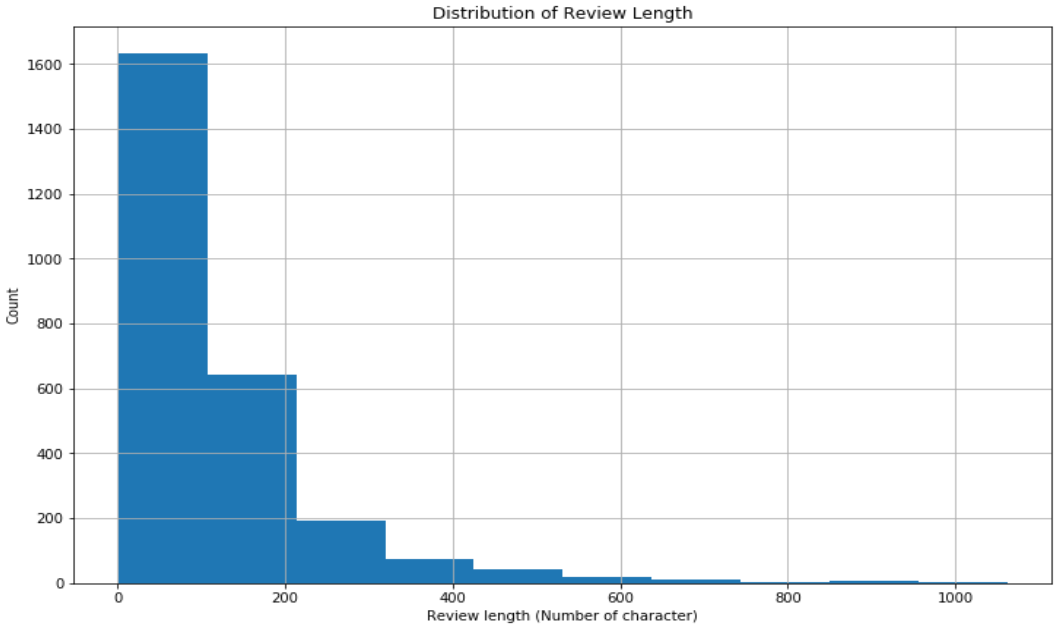
1540 and 223 reviews of values five "5" and four "4" respectively in the Figure (4) form the positive count in the corpus, while 167 and 693 reviews correspondingly of values two "2" and one "1" form the negative count reviews. The remaining class of value three "3" stands for neutral sentiment with 12 reviews count.

***Figure 4**- Distribution of rating count*



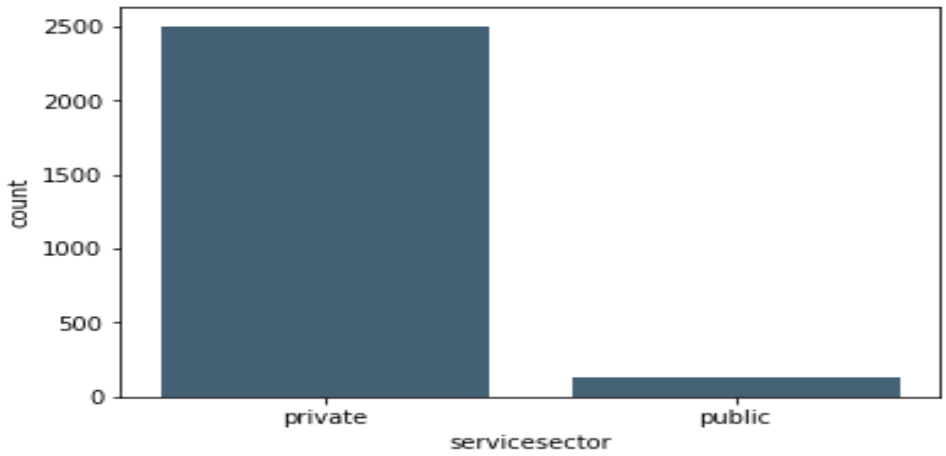
Text reviews show distinctive distributions in terms of character length. More than half of the reviews of approximately 1620 are between 0 to 100-character length. 620 reviews have a length between 100 to 210-character length. Afterwards, the reviews count decreased to 198 with length in-between $210 - 310$ characters. Reviews count continues to decrease to reach of about 3 with length of $950 - 1050$ characters. Overall, with the increase of review length in x-axis there is an associated decrease in review count representation in the y-axis.

***Figure 5**- Distribution of Review Length Across The Arabizi Corpus*



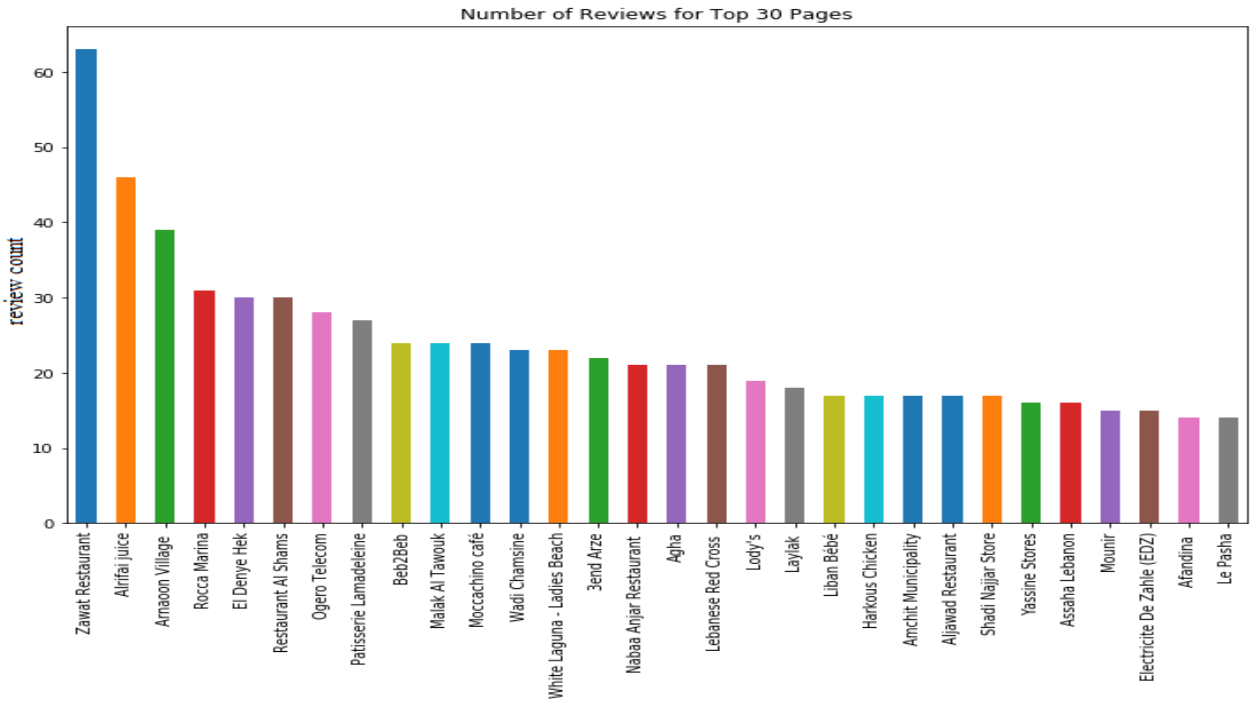
Moreover, each review in the dataset is assigned to the service sector that belongs to. There are two distinctive service sectors that are the public governmental institutions and the other is specified for the private institutions, shops, restaurants, hotels, etc. The total reviews, which belongs to the private service sector, are 2501. On the other hand, the total public service sector reviews are 134.

***Figure 6**- Distribution of Review in Respect to Service Sector Providers: Private & Public*



Furthermore, the reviews in the Arabizi corpus is extracted from many pages in the three different websites: Facebook, Google, and Zomato. The total number of page names are 776, which is sample representative. The top 30-page names that contain the highest review count is specified in the following figure:

***Figure 7**- The Top 30-page names that contain the highest review count*



3.2.1 **The Challenges of Analyzing Arabizi Texts**

Arabizi is very common-used language system within the new generation in the Arab world, claiming that it is easier and much faster than texting in Arabic, which is not friendly with technology (Basis Technology, 2012). In addition, the Arabizi is a growing challenge for institutions especially the government intelligence agencies, for it does not follow a systematic language system. It is expressed and written differently based on the web user's experience. In the following subsections, we presented few linguistic challenges that occur frequently in the collected corpus of Arabizi texts. Because of its highly importance, we would present the following challenges that we have found while analyzing data for the favor of building the SLCSAS classifier.

3.2.1.1 ***Exaggerations***

Internet users tend to exaggerate while expressing their opinions in social media. Consider the following examples:

Figure 8*- Exaggeration in Arabizi texts*

Arabizi texts	**English equivalent**
Jmiiiil jmiiil jmiiil	Beautiful beautiful beautiful
Raw333333a	Amazing
Ktyyyyyyyyyyyyyyyyyyyyyyyyyyyyyr helwin	They are very pretty
Wawwww	Wow
ktiiiiir helwe	She is very pretty
ktir ktir ktir jorsa stuff	Very very very waste stuff
Ktirrrrr Tayib	Very delicious

From the above examples, we could notice that exaggeration takes two distinct shapes. The first is on the word level in which the word contains repetitive letters as in Wawwww/wow, while the other is based on repeating the whole word more than one time as in ktir ktir/very very.

3.2.1.2 ***Code Switching and Mixing***

Code mixing, or switching is the alternating use by bilingual language users of two or more than languages within an utterance (Muysken, 2000).

Table 7*- Code Mixing and Switching Phenomenon in Lebanese Arabizi Corpus*

Arabizi texts	**English equivalent**
We found 2 pieces of chiken fo2 ba3ed	We found 2 pieces of chicken over each other

I'm an everyday customer Mberha brou7 la e7dar	I am an everyday customer, yesterday I went to attend
؟؟طريق عام ??shou ya3ne Explain to me why valet parking should take money and i asked el baladiye main road??	What does that mean?? Explain to me why valet parking should take money and I asked the municipality.
2akal mnel 3ade... WORST EXPERIENCE EVER	food from normal… worst experience ever

3.2.1.3 ***Question Sentences***

Some sentences have either positive or negative polarity but if it is a question, it may have a neutral polarity.

Table 8*- Examples of Question Arabizi Sentences*

Arabizi texts	**English equivalent**
Wen natiji natrin 3a naaaar	Where is the grade, we are waiting on fire
3endak shi need for speed payback lal ps3 ma3 delivery	Do you have Need for Speed Payback for PS3 and with delivery
W3n bi ser haydah ?	where is this thing?
Kifak	How are you?
Iza baddi ammel sneni hek. Addesh bikalefni	if I would like to do my teeth like his. How much it costs me?

3.2.1.4 ***Opposite Representation (Intensifiers)***

This type of sentences is based on linguistic challenges, which have negative lexical polarity.

Table 9*- Opposite Representation of Negative Lexical Markers*

Arabizi texts	**English equivalent**

ma atybak	How delicious
w5edmat ma ba3d	and service after (purchase)
btekol ad ma tayib w ndif	(you) eat as much as it is delicious and clean

The first and last "ma" is not a negative word but, in this case, it is an intensifier for the adjective atybak and tayib/delicious. The second one in ma ba3d/and after is not a negative marker but a transitional word equivalent to after in English language.

3.2.1.5 ***Part of Speech Tagging (POST)***

POST is an automatic or a manual assignment one of the parts of speech to a word. This task includes nouns, verbs, adverbs, adjectives, pronouns, conjunction and other sub-categories (Oudah and Shaalan, 2012). POST aids ALP and IR tasks.

Ktir 7eloo/very beautiful → Ktir: adverb,

7eloo: adjective.

The adjective "7eloo"/"beautiful" gives description for the noun it follows or precedes in a sentence as like "ma7al"/"shop". The adverb, however, changes in position to add extra and exaggeration to the adjective and noun it describes as in "el ma7al ktir 7eloo"/"the shop is very beautiful" and can be "el ma7al 7eloo ktir"/"the shop is very beautiful".

3.2.1.6 ***Affixes***

In Modern Standard Arabic (MSA), words change its contextual, semantical, syntaxial and morphological linguistics components based on affixes. Consider the following examples in changing words from singular adjective to plural, affirmative noun, and plural verb, as follows:

Table 10*- Variations in Linguistics components of Arabizi words*

Singular adjective	Plural adjective	Affirmative noun	Plural verb
nassab	nasaben	lal nasb	ynasboo

7arameye	7arammiyyi	…	…
ta2ifi	ta2ifyin	…	…
…	kizzabeen	kezeb	…

In addition, plenty of terms in Arabizi system concatenated most often with the articles that proceeded, as follows:

***Table 11**- Arabizi Texts Concatenation Phenomenon*

Arabizi texts	**English equivalent**
se3a latousal	An hour to be received
l5dme	The service
lwaiters	The waiters
l5adamet	The services
lshabeb	The young

3.2.1.7 ***Superlative and Comparative Adjectives***

Arabic language system is rich in morphological and syntactical components. Like the Arabic language properties, Arabizi word has a lot of conjugation forms. Comparative adjectives compare two things using the term plus a word that is in our case mn/from, whereas superlative adjectives compare one to thing to group of things.

***Table 12**- Superlative and Comparative of Arabizi Adjectives*

Adjective	**English equivalent**	**Comparative**	**Superlative**
7eloo	nice	a7la mn	a7la
kezeb	lie	akzab mn	akzab

kbeer	big	akbar mn	akbar
tyeb	delicious	atyab mn	atyab

3.2.1.8 ***Writing Variations***

Too many words with writing variations do exist in the Lebanese Arabizi corpus and that goes back to the different user's educational background and experience exposure, a sample looks as follows:

Table 13*- Arabizi Terms Variations in Writing*

Arabizi texts	**English equivalent**
a7la, ahla, 27la	Most beautiful
ktir, kter, ktr	Very
kelshi, klchiii, klshi, kl, chi	Everything
7lwe, 7alw, helwh	Beautiful

3.3 **Data Preprocessing and Filtering**

Data have been preprocessed in attempt of sharpening the overall accuracy of classifiers. The following summarizes all the steps that data have been applied to.

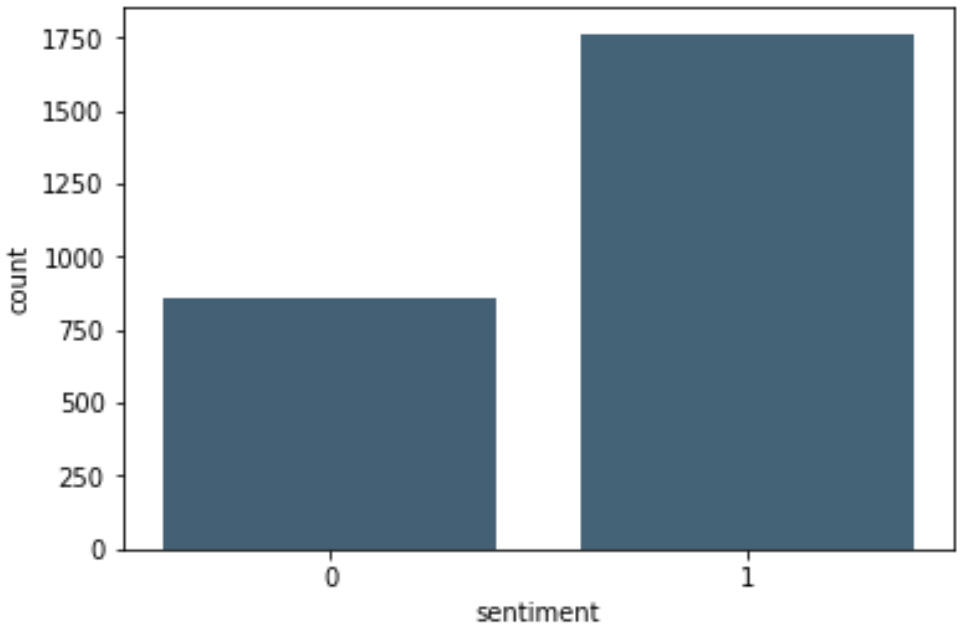
3.3.1 **Removal of reviews with "neutral" sentiment**

Due to neutral class smallest representation in the dataset, reviews with value of "3" dropped from the target dataset to avoid confusion in classification. If its representation is equal to at least the negative sentiment bar, we would keep it for representing the neutral sentiment class. However, it represents 0.46% of the dataset with 12 reviews count.

3.3.2 **Ratings' Encodings**

Because we are dealing with a binary sentiment classification problem, reviews with values of $5s$ and $4s$ encoded as 1 (representing the positive sentiment); reviews with values of $1s$ and $2s$ encoded as 0 (representing the negative sentiment). The final representation of both sentiment classes in the x-axis and the sentiment classes' counts in the y-axis, which looks as follows:

***Figure 9**- Sentiment Classes Distributions in Arabizi Corpus*



The positive class of value "1" remarks 1755 reviews count, while the negative class of value "0" is 880 texts count.

3.3.3 **Data splitting for training and testing**

Data splitting step is crucial for ensuring unbiased reviews' selections in research experimentation. To have the data ready for classifier's parsing and to maintain equal comparison between the two taken approaches of the ML and the rule-based one, the dataset split randomly 80% for training and 20% for testing. The training set is 2098 examples with 684 negative and 1414 positive reviews. On the other hand, the testing set is 784 examples with 176 negative and 349 positive reviews.

3.3.4 **Data Cleaning**

Data cleaning is important for later stages of feature extraction. Cleaning reviews are easier to process and manipulation, which includes the following:

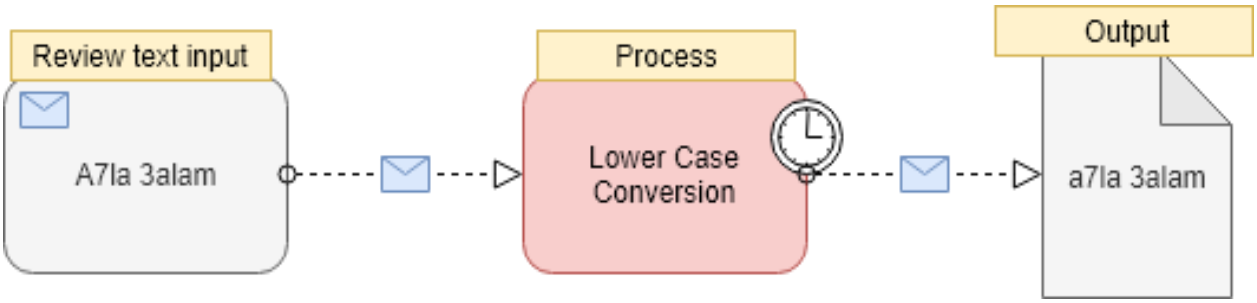
3.3.4.1 ***Removal of User's Related Information***

To ensure maintenance of ethical standards, all private information related to the user is deleted, as such user name and posting time period.

3.3.4.2 ***Lower Case Conversion***

Each letter of a word that is in a review converted to a lower case for the use of words entity matching. This step is very useful and demanding to avoid overlapping of same terms but with different case representations.

Figure 10*- Example of Text Review a7la 3alam/The best people Conversion to Lower Case*



3.4 **Reviews Representation**

Because ML algorithms work with numeric data, the provided corpus data could not be used as it is. Therefore, BoW, and TF*IDF used for this purpose, which is discussed with details in the following subsection. On the other hand, this step is ignored in the adopted rule-based approach.

3.4.1 **Selected Features**

To conduct a supervised learning for OM, there is a sequence of steps to follow as converting the text to numeric data presented in vector space, which eases the data mapping with labels, also efficiently performing feature extraction and selection to train the used classifier on

the dataset. Then, we could estimate the error based on the test dataset much quickly and more efficiently.

3.4.1.1 ***Bag of Words (BoW) model***

According to Manning, Raghavan, and Schütze (2009, p. 117), BoW is intuitively a simple model to represent terms' occurrences for a given document in comparison to other models as TF*IDF. For BoW creation, each term t in documents d would be regarded by assigning a weight to t based on the frequency of occurrences of t in d. This method called term frequency (tf_t, d). For this thesis, documents d represents reviews in the corpus and t is the frequency of occurrences of terms in reviews in word levels n-grams (unigrams, bigrams, and trigrams) after they have been passed through the preprocessing step. Thus, the following reviews $\boldsymbol{d_1}$ ("a7la mat3m a7la 3alam"), $\boldsymbol{d_2}$ ("A7la café"), and $\boldsymbol{d_3}$ ("A7la mat3m") are identical by same terms' occurrences, which correspond to the following BoW representation:

Table 14*- BoW: Terms Frequency of three sample reviews*

$\boldsymbol{t/d}$	$\boldsymbol{d_1}$	$\boldsymbol{d_2}$	$\boldsymbol{d_3}$
a7la	2	1	1
mat3m	1	0	1
3alam	1	0	0
café	0	1	0
a7la café	0	1	0
a7la mat3m	1	0	1
mat3m a7la	1	0	0
a7la 3alam	1	0	0
a7la mat3m a7la	1	0	0
mat3m a7la 3alam	1	0	0

However, this model shows a great weakness through assigning a weight to every t of a given d. For example, stopwords remark a high frequency rate, but they do not contain any valuable idea.

3.4.1.2 ***Term Frequency and Inverse Document Frequency (TF*IDF)***

To spot the meaningful terms unlike BoW model, the TF*IDF mechanism would be an excellent fit for representing term t in given documents d. Firstly, the idf mechanism recognizes less frequent terms in given documents by assigning them a large weight, while assigning a low weight for most frequent terms in these documents (Manning, Raghavan, and Schütze, 2009, p.118), by the following formula:

$$idf_t = log \frac{\#df}{1+\#df_t}, \quad (2)$$

where the inverse document frequency idf of a term t is equal to logarithmic of the total number of documents over the documents that contain that term t plus one to avoid sparse denominator.

Table 15*- Inverse Document Frequency of Three Sample Reviews*

t/idf_t	idf_t
a7la	-0.124
mat3m	0
3alam	0.176
café	0.176
a7la café	0.176
a7la mat3m	0
mat3m a7la	0.176
a7la 3alam	0.176
a7la mat3m a7la	0.176

mat3m a7la 3alam	0.176

In addition, tf is basically the output of the BoW model, which determines how important a term t is by looking at how frequently it appears in the documents d. If a word appears a lot of times, then the term must be important. Therefore, the TF*IDF model is distinguished from BoW model by the following formula:

$$tf * idf, d = tf_t, d \times idf_t \quad (3)$$

This feature model was used for it distinguishes terms that are seen in small number of documents and offers a less pronounced relevance signal for those appearing in all documents.

***Table 16**- TF*IDF of Three Sample Reviews*

t/d	d_1	d_2	d_3
a7la	-0.248	-0.124	-0.124
mat3m	0	0	0
3alam	0.176	0	0
café	0	0.176	0
a7la café	0	0.176	0
a7la mat3m	0	0	0
mat3m a7la	0.176	0	0
a7la 3alam	0.176	0	0
a7la mat3m a7la	0.176	0	0
mat3m a7la 3alam	0.176	0	0

3.5 **Research Tools**

The classifiers that are used in this study in SA. Firstly, it includes LR model. In addition, another classifier based on hand-created rules were used under the name SLCSAS. Both classifiers were discussed in detail in the following parts.

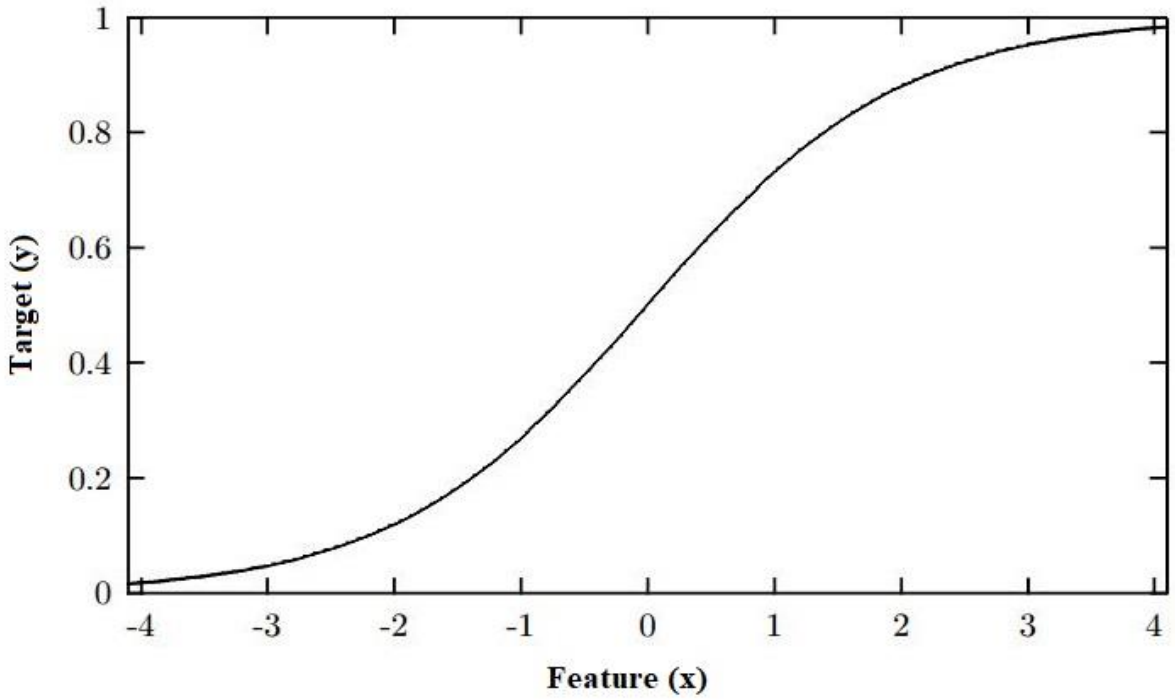
3.5.1 **Machine Learning Classifier**

In this section, LR algorithm would be discussed in favor of SA task, providing a concrete example besides the theoretical background at its utmost simplification.

3.5.1.1 ***Logistic Regression (LR)***

LR model uses logistic function as its core, which is also called sigmoid function. This function presented by a S-shaped curve, which takes a number between $+\infty$ and $-\infty$ to map it into a value scale between 0 and 1.

Figure 11*- Logistic (sigmoid) Function specifying the feature x of input x_i and the target y sentiment class, adopted from Cramer (2002, p.3)*



The function drawn in Figure (11) represents the generic function before mapping corresponding results into the logistic function. This function interpreted as the score sentiment $Score$ of given input x_i is equal to the learnt coefficient of features found in that given input x_i. This function can be simplified into the following:

$$Score(x_i) = w_0 h_0(x_0) + w_1 h_1(x_1) + \cdots + w_n h_n(x_i) \quad (4)$$

To illustrate this function, a one should consider the following example. Supposedly, there is an input x_i that said, "Sushi was great, the food was awesome, but the service was terrible," and thus the trained features of that given x_i is coefficient value of trained features, for example, the frequency of awesome and terrible. And, the coefficients of these awesome and terrible is given by the following table:

Table 17*- An Example illustrating Score (x_i), Representing Value Coefficients of a Given Input (x_i)*

Input (x_i)	**Coefficient (w_i)**	**Value**
Intercept (default)	w_0	1.0
#awesome	w_1	1.0
#terrible	w_2	-1.5

Accordingly, the $score$ of the input text "the food was awesome, but the service was terrible" is:

$$Score\ (x = \text{"}the\ food\ was\ awesome, but\ the\ service\ was\ terrible\text{"}) = 1.0\ \#awesome - 1.5\ \#\ terrible = -0.5, \quad (5)$$

where the answer is close negative, which corresponds to the negative sentiment class if it is considered in logistic function.

From the logistic function introduced above, we could calculate the LR by taking the sigmoid of the computed $Score$ of input x_i.

$$P\ (y = \pm 1|\ x_i, w)\ =\ sigmoid\ (Score(x_i)) \quad (6)$$

Equation (6) corresponds to the LR function, which means that the $P\ (y = \pm 1|\ x_i, w)$ is the probability (P) of estimated output (y) having said that represented by vertical line (|), input (x_i) and coefficient values (w) that is current in x_i equal to the sigmoid of ($Score(x_i)$) that alternatively could be represented by the following equation:

$$p(y = \pm 1|x_i, w) = \frac{1}{1+e^{-w^{(t)}T*h(x_i)}}, \quad (7)$$

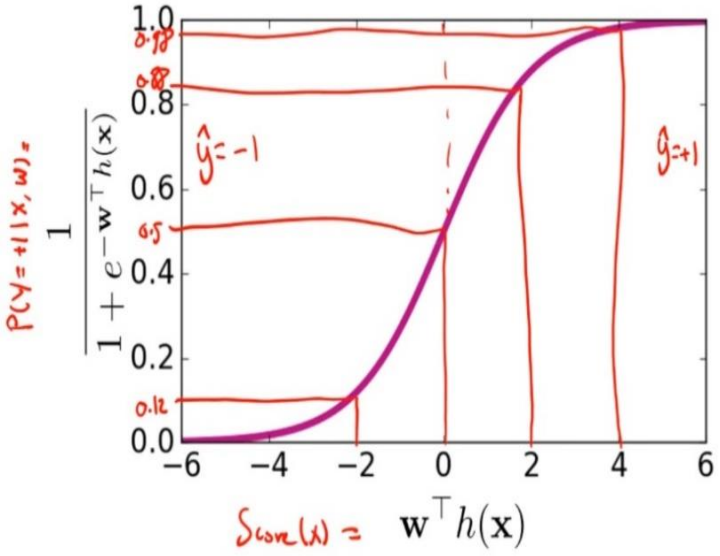
where the probability of the review to be equal to positive class ($+1$) or negative class (-1) based on the given input (x_i) given features learnt coefficient (w). The function computed by 1 divided by 1 plus exponential (e) powered to the dot product of transpose learnt coefficient (w) in iteration t of each feature (h) in the given review input (x_i).

The following presents the squeezing of sigmoid function for sentiment probability in which "0" represents negative class, "$+1$" represents the positive class, while "0.5" represents the uncertainty neither of both sentiments.

$$y = \begin{cases} 0, & x < 0 \\ 0.5, & x = 0 \\ +1, & x \geq 0 \end{cases} \quad (8)$$

An example is to be understanding the LR model much better. Let's just take two $scores$ of x_i 0 and 2. From the graph, which represents the S-shaped LR model, the $score$ of 0 if it is applied to it, the corresponds probability would be 0.5 which is neither positive nor negative sentiments. In addition, the $score$ of 2 has the probability of 0.88, which is obviously a positive sentiment class.

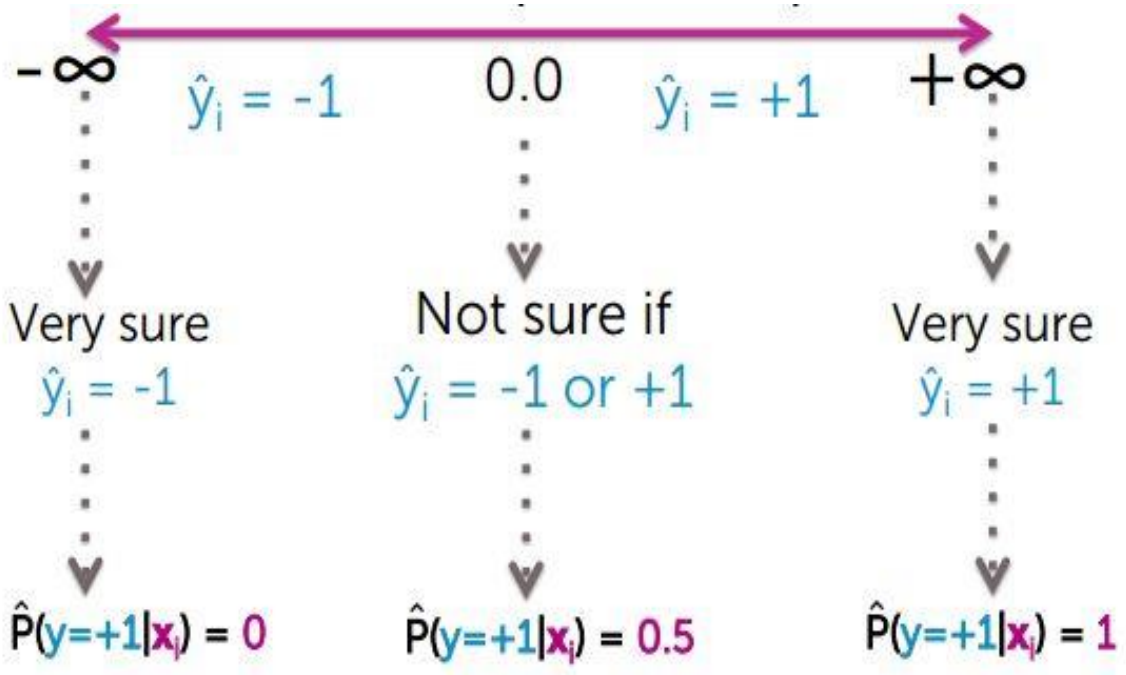
***Figure 12**- An Example of OM using LR, adopted from (Guestrin & Fox, 2016a)*



Score(x_i)	P(y=+1|x_i, w)
0	0.5
-2	0.12
2	0.88
4	0.98

Furthermore, concerning the output probability of any classifier in this case is the LR, the probability can be further illustrated to navigate classifier's confident in the task of SA. From the example provided in Figure (12), the $score$ of 0 is equal to 0.5, which could signify the unsureness of giving neither a positive nor negative sentiments' classes. The following figure gives a glimpse of classifier confident:

***Figure 13**- Classifier Confident in the Task of OM, adopted from (Guestrin & Fox, 2016b)*



Furthermore, LR uses measures quality of fit for model with coefficients w. The maximum likelihood estimation (MLE) finds the best classifier with respect to line s-shaped curve over all possible coefficients through computing the sum of probabilities of both classes of negative and positive reviews. The following formula carries the computation of the maximum likelihood over all data points N with regard to each feature x_i to maximize coefficient w, and therefore picking $\widehat{w}$ that makes this function as large as possible.

$$l(\widehat{w}) = \prod_{i=1}^{N} P(y_i|x_i, w) \quad (9)$$

A quick example, adopted from Guestrin and Fox (2016c), would be as follows in which we have four data points that have feature x_i the frequency of awesome and awful that we want to maximize the coefficient w to as near as to $+1$ for positive examples and 0 for negative ones. Therefore, LR model would fit a curve to fit the data in binary classification space in which data with positive class would be pushed to each other in the space and data in the negative class would

be pushed to the minimal coefficient representation. The following table represents the MLE parameters maximization for finding the best $l(\widehat{w})$ as in the following table:

***Table 18**- Example: Coefficient Maximization*

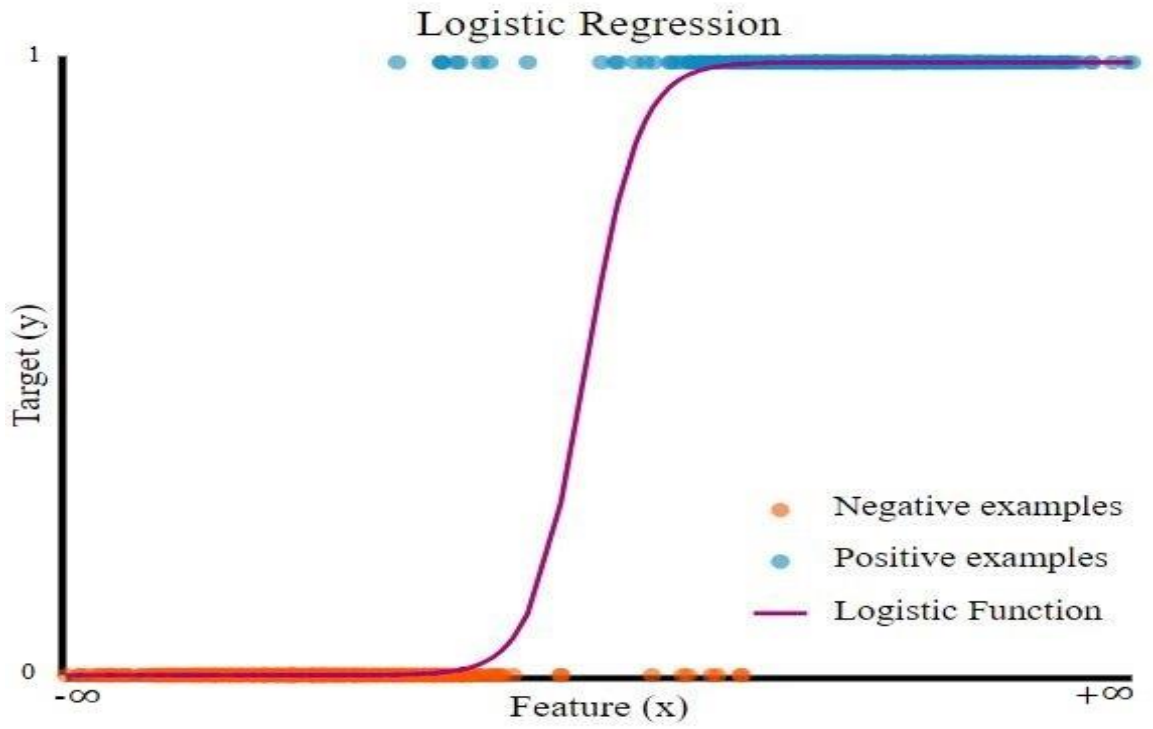
Data points	x_1	x_2	y	**Choose w to maximize**
x_1, y_1	2	1	$+1$	$P(y=+1\|x[1]=2, x[2]=1, w)$
x_2, y_2	0	2	-1	$P(y=-1\|x[1]=0, x[2]=2, w)$
x_3, y_3	3	3	-1	$P(y=-1\|x[1]=3, x[2]=3, w)$
x_4, y_4	4	1	$+1$	$P(y=+1\|x[1]=4, x[2]=1, w)$

Therefore, the MLE is equal to the multiplications of probabilities of data points, which would be presented for the previous example in the following computation of equation (9).

$$l(\widehat{w}) = P(y=+1|x[1]=2, x[2]=1, w) * P(y=-1|x[1]=0, x[2]=2, w) * P(y=-1|x[1]=3, x[2]=3, w) * P(y=+1|x[1]=4, x[2]=1, w) , \quad (10)$$

For clarification, MLE works on maximizing reviews with positive class to 1 and negative ones to 0. Therefore, the LR line would look like as in the following plot in which we can notice most of data points with positive instances are:

***Figure 14**- Data Representation in LR on OM Task*



In addition, the LR model computes the derivate of each coefficient feature with respect to all other candidate features. The derivate is computed through taking partial derivate of coefficient parameter to do the update with respect to iteration step ŋ.

$$\frac{\partial l(w^{(t)})}{\partial w_j} = \sum_{i=1}^{N} h_j(x_i)(\mathbb{1}[y_i = +1] - P(y = +1|x_i, w^{(t)})), \quad (11)$$

where the partial likelihood coefficient w in iteration t with respect to w_j is equal to the computation sum of all the data points in which the feature in input review x_i is multiplied by the true predication minus the predicated probability by the trained LR with all respect to current

coefficient $w^{(t)}$ in iteration t. The following example simplifies the process of coefficient learning and update in LR.

***Figure 15**- Example: Coefficient values*

Coefficient	value
$w_0^{(t)}$	0
$w_1^{(t)}$	1
$w_2^{(t)}$	-2

***Figure 16**- Example: Computation of Derivate Contribution to Coefficient w_1*

$x[1]$	$x[2]$	y	$P(y=+1\|x_i,w)$	Contribution to derivate for w_1
2	1	$+1$	0.5	$2(1-0.5)=1$
0	2	-1	0.02	$0(0-0.02)=0$
3	3	-1	0.05	$3(0-0.05)=-0.15$
4	1	$+1$	0.88	$4(1-0.88)=0.48$

The total derivate of w_1 would equal to the following:

$$\frac{\partial l(w^{(t)})}{\partial w_1} = 1 + 0 - 0.15 + 0.48 = 1.33 \quad (12)$$

Therefore, the update in iteration t to reach to the function's convergence is computed through taking the coefficient of chosen parameter $w_1^{(t)}$ plus the step size multiplied by the partial likelihood coefficient w in iteration t with respect to w_j, which was already obtained in calculation (12). Accordingly, as we see from the following formula, the result is 1.133 that is updating the initial value of $w_1^{(t)}$ from 1 to 1.133. This process goes over all coefficients (parameters) in a

number of iterations that preselects till the norm derivate of $l(w^t)$ is equal to 0, however we set a threshold tolerance that the function converges whenever it reaches to.

$$w_1^{(t+1)} = w_1^{(t)} + \eta \frac{\partial l(w^{(t)})}{\partial w_1} = 1 + 0.1 * 1.33 = 1.133 \quad (13)$$

Finally, it must be noted that regularization adopted in this master's project is the L2 regularization norm. In L2 norm, the coefficients are computed through sum of their squares to avoid sparse coefficients as in L1 norm that sums the absolute values. The lambda parameter is subject for tuning for penalizing w coefficients. The larger λ is less likely the coefficients will be increased in magnitude to adjust for small perturbations in the data.

$$\lambda \parallel w \parallel_2^2 = w_0^2 + w_1^2 + \cdots + w_j^2 \quad (14)$$

3.5.2 **Lexicon-based Classifier**

In this section, the SLCSAS algorithm would be discussed in favor of NLP tasks, providing description of program's uses. Finally, we would represent an example for quick demonstration.

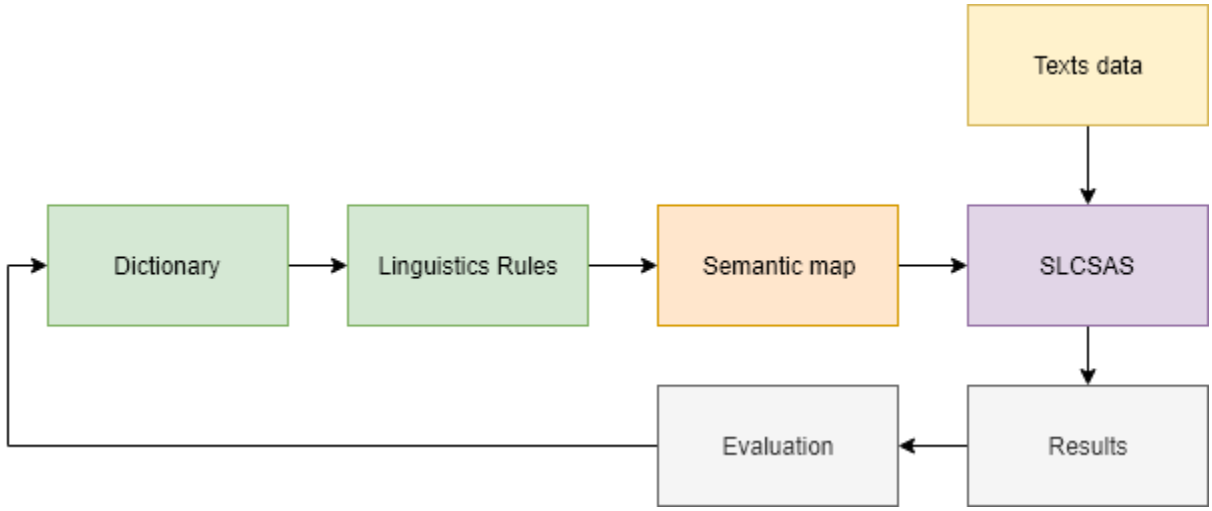
3.5.2.1 ***SLCSAS***

SLCSAS (Science of Language and Communication Semantic Analysis System) a program written in Perl programming language developed at the Lebanese University, Centre for Language Sciences and Communication, Celine Centre, Tayouneh, Beirut, Lebanon by Moustafa Al-Hajj (2018). The program allows computational linguists researchers to create their own sophisticated linguistic grammar and dictionary that are useful for the classification task following a semantic map that specifies the grammar and dictionary categories.

In general, SLCSAS handles corpora that contain natural language data in txt format. It is useful for wide range of tasks including IR, part of speech tagging (POST), speech act classification, name entity recognition (NER), OM, and many others (Al-Hajj, 2018). SLCSAS classifier is based on a dictionary of terms following a collection of linguistics rules belonging to defined semantic categories. The rules and grammars are hand-crafted ones, which are up to the linguist own analysis on the field of research. Once they have been built the dictionary, grammar, and the semantic map (semantic categories), SLCSAS classifier processes the linguistics rules that distinguished in the semantic map on the input texts data. The results would be outputted in html

format with the hand-crafted rules marked with yellow color to validate whether it matches what is intended to or not. Then, evaluation is conducted according to the obtained results. Finally, adding new terms to the dictionary is possible to sharpen the classifier's performance. It is also possible to add new rules and make the necessary edits to the semantic map.

***Figure 17**- SLCSAS Architecture*

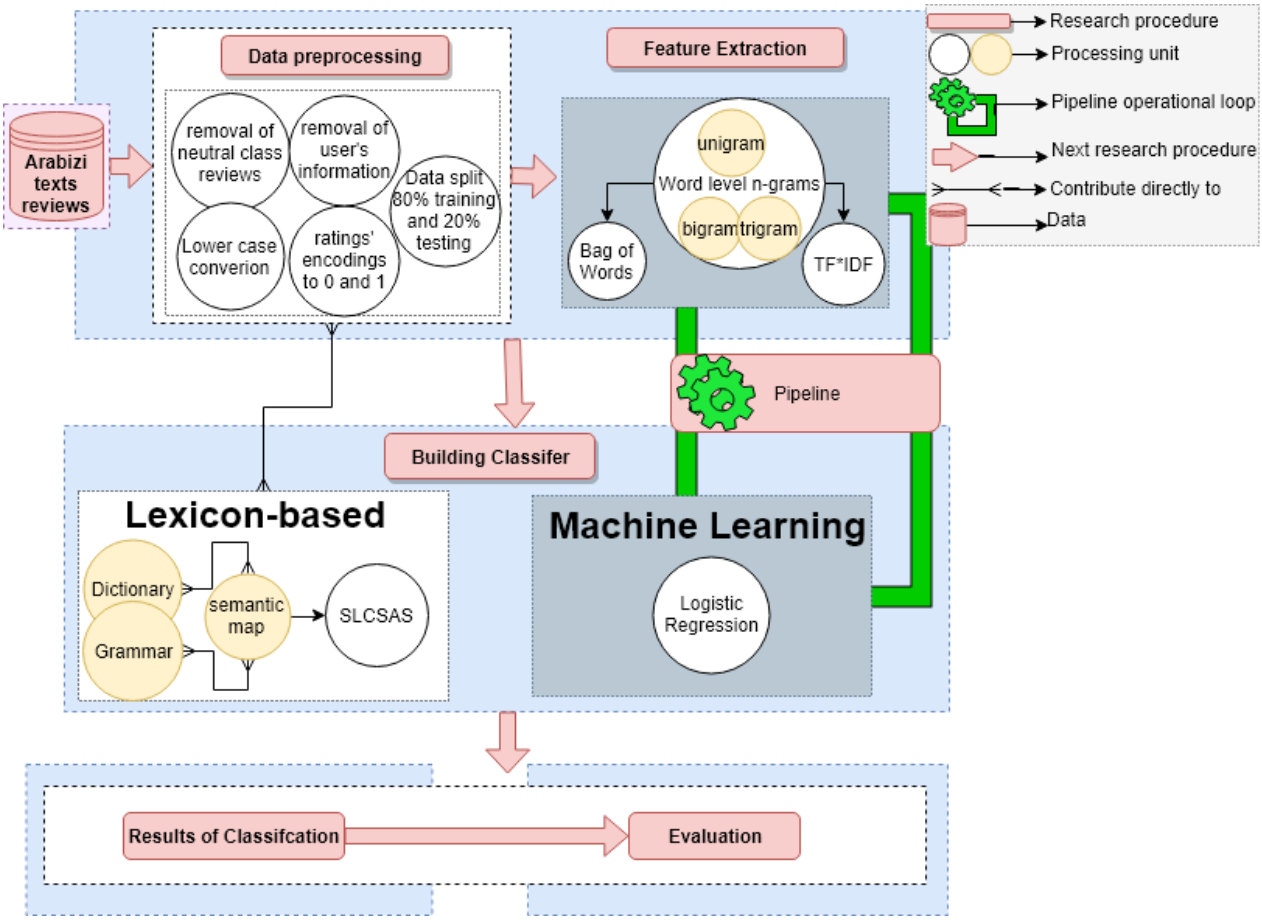


In this thesis, linguistics markers with target sentiment texts approach were used to build the SA SLCSAS classifier with wide semantic categories, which will be specified later in **IV. Experiment Preparation** section.

3.6 **Research Procedure**

To proceed with the SA task, Arabizi text reviews are being feed into steps of filtering and preprocessing that was intuitively discussed earlier in 3.3 **Data Preprocessing and Filtering** to having a readily polished data for the task of SLCSAS and the ML training and testing. For ML implementation, word level n-grams (unigrams, bigrams, and trigrams) TF*IDF and BoW features extracted from data in attempt to achieve the highest score if possible. In addition, a pipeline loop invested in hyperparameter tuning of the LR model. On the other hand, lexicon-based SLCSAS classifier's dictionary, grammar, and semantic map have been built manually through deep analysis of preprocessed data. Then, building the classifiers with corresponding extracted features from data is necessary as it is specified in Figure (18) below. Finally, classification results are computed for the final evaluation of each undertaken classifier.

***Figure 18**- Research Procedure to Approach SA*



IV. Experiment Preparation

In the previous sections, we have described the necessary literature background regarding SA, applicable approaches through previous research studies, data feature extraction, preprocessing techniques, and finally LR and SLCSAS classifiers. In this section, we will describe each research procedure presented in Figure (18) above. First, data preparation that includes data preprocessing and feature extraction of Arabizi text reviews in the Lebanese dialect that required for conducting experiments of both ML and lexicon-based classifiers. Then, rule-based SCLSAS main components of dictionary, grammar, and the semantic map will be discussed in detail as well

as the LR ML model will be discussed with the pipeline *gridsearch*[7] approach. At the end, we will present what are the measure we will take to assess the performance of each classifier.

4.1 **Data Preprocessing**

Each review corresponds to a reviewer and a rating in the dataset, which was contained in .csv file (A Comma Separated Values UTF-8), is so necessary to chop off all reviewers' names from the dataset to maintain ethical standards, and to encode ratings from scale 5 to 1 into a scale from zero to one because we are dealing with binary classification problem. Reviews with neutral sentiment class of 3 dropped from the dataset. For the experiments' evaluations, we need to separate our dataset into training and testing sets. For this purpose, we randomly split the dataset into training and testing subsets (80% and 20%) using the free *scikit-learn* library (*sklearn.model_selection.train_test_split*)[8], and later both testing and training subsets are extracted in .txt files to apply both approaches of ML and rule-based on the same data split . For hyperparameters tuning in ML model that is LR, we applied the n-fold cross-validation procedure. In a n-fold cross-validation the classifier is trained on $n - 1$ folds of the data and tested on the remaining folds, then this process is repeated n times for different splits, and the results are averaged over the n experiments. We used 2 folds of cross-validation on the training set because of the small sample size. Besides, all text reviews are lower cased for terms' matchings. Therefore, these are the only steps we took to preprocess the data because we believe that every letter matters in the dataset. In this way, the classifiers would meet the real-world classification problem with all it enfolds within texts in SA task.

4.2 **Feature Extraction**

For ML, We used the vector space models (VSMs) for representations of texts, including BoW and TF*IDF features with word level n-grams (unigrams, bigrams, and trigrams). They have been used to train the LR ML model. Table (19) lists, for the dataset, the total number of the extracted features for each of TF*IDF and BoW unigram, bigram, and trigram that are used for LR model training.

---

[7] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[8] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

***Table 19**- TF*IDF and BoW word level n-gram features*

Unigram	Bigram	Trigram	Total
12,637	32,580	34,802	80,019

In BoW, reviews are presented by a table in which the columns represent the existing words in each review and the values represent their frequencies. Therefore, a collection of reviews can be represented as illustrated in Table (20) in which there are n reviews and m lexicons. Each review is represented as follows:

$$review_{it} = (w_{i1}, w_{i2}, \dots, w_{im}), \quad (15)$$

where w_{ij} is the word's frequency (or n-words in case of using n-grams) w_{ij} in the $review_{it}$. *scikit-learn* library *CountVectorizer*[9] used to build the vector model of unigram, bigram, and trigram BoW features for the dataset.

***Table 20**- BoW Feature Representation Matrix*

$review_{it}$	w_1	w_2	…	w_m
review$_1$	$freq_{11}$	$freq_{12}$	…	$freq_{1m}$
review$_2$	$freq_{21}$	$freq_{22}$	…	$freq_{2m}$
…	…	…	…	…
review$_n$	$freq_{n1}$	$freq_{n2}$	…	$freq_{nm}$

On the other hand, TF*IDF presents reviews by a table in which the columns represent the existing words in each review and the values associated to the multiplication of existing word frequency (globally) by inverse document frequency (rarity). Thus, a collection of reviews can be

[9] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

represented as illustrated in Table (21) in which there are n reviews and m lexicons. Each review is represented as follows:

$$review_{it}^{tfidf} = (w_{i1}^{tfidf}, w_{i2}^{tfidf}, \dots, w_{im}^{tfidf}), \quad (16)$$

where w_{ij} is the existing word frequency in the taken $review_{it}$ multiplied by its uniqueness in all the reviews, using also the *scikit-learn* library *TfidfVectorizer*[10] used to build the vector model of unigram, bigram, and trigram BoW features for the dataset.

***Table 21**- TF*IDF Feature Representation Matrix*

$review_{it}^{tfidf}$	**w1**	**w2**	**…**	**wm**
review1	$tfidf_{11}$	$tfidf_{12}$	…	f $tfidf_{1m}$
review2	$tfidf_{21}$	$tfidf_{22}$	…	$tfidf_{2m}$
…	…	…	…	…
reviewn	$tfidf_{n1}$	$tfidf_{n2}$	…	$tfidf_{nm}$

4.3 **Building Classifiers**

Two distinctive classifiers were used for Arabizi SA task. The first is the LR ML model, and the other is SLCSAS. Both will be built following a procedure approach, which will be discussed in detail. Both approaches of ML and lexicon-based on a dataset of 2635 Lebanese Arabizi reviews. LR and SLCSAS classifiers tested both against the best performance achieved on the dataset. The ML approach has two main training phases on 80% of the dataset (2098 reviews) and testing on 20% (525 reviews). At the first phase, two experiments conducted using ML model: LR paired with BoW, TF*IDF, and world level n-grams features. At the second phase, another two experiments carry hyperparameter tuning for LR models through Pipeline architecture. In the third experiment, SLCSAS rule and dictionary classifier-based tested on 20% of the dataset (525 reviews) after the dictionary is constructed through deep analysis of Arabizi text reviews to extract

[10] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

keyword markers and sentiment texts for positive and negative classes on 25% of the training set (525 reviews).

4.3.1 **Machine Learning**

For better performance on the SA Arabizi task, we used the free *scikit-learn* library of LR[11]. Accordingly, LR models trained in two separate distinctive training phases. The first phase, two LR models were trained on default parameter settings. The second phase includes hyperparameter tuning for the purpose of hacking the performance of LR model.

4.3.1.1 ***First phase (Default settings)***

At the first phase in the first experiment, two LR classifiers with BoW and TF*IDF and n-grams features without hypermeters tuning trained on 80% of the dataset and tested on 20%. The first model of LR trained on the 80% of dataset through enhancing its base vocabulary by using TF*IDF with word levels of unigrams, bigrams, and trigrams altogether. On the other hand, the second model of LR trained on the dataset with BoW that contain different word levels of unigrams, bigrams, and trigrams altogether. The architecture of both train models includes the following main specifications where n-gram range is in unigram, bigram, and trigram, 1 regularize strength, 0.0001 tolerance for function's convergence, and with 100 maximum iterations updates.

Table 22*- Main Default LR settings*

Parameter\architecture	**BoW LR**	**TF*IDF LR**
multi-level n-gram	unigram, bigram, and trigram (1,3)	
inverse of regularization strength	1	
tolerance	0.0001	
maximum iterations	100	
Penalty	L2	

[11] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

The remaining default architecture models' settings is extracted from the built model in scikit-learn, which specifies the following:

LogisticRegression (C=1.0, class_weight=None, dual=False, fit_intercept=True, intercept_scaling=1, max_iter=100, multi_class='warn', n_jobs=None, penalty='l2', random_state=None, solver='warn', tol=0.0001, verbose=0, warm_start=False)

The following Table (23) represents an example of 6 top BoW and TF*IDF features with smallest and largest coefficients.

***Table 23**- Learnt Coefficients of both LR Models with BoW and TF*IDF features*

BoW LR		**TF*IDF LR**	
Words with smallest Coefficients	**Words with largest Coefficients**	**Words with smallest Coefficients**	**Words with largest Coefficients**
tfeh (spit)	a7la (most beautiful)	ma (not)	a7la (most beautiful)
ma (not)	atyab (most delicious)	3al (on)	atyab (most delicious)
zbele (rubbish)	raw3a (magnificence)	tfeh (spit)	raw3a (magnificence)
se3a (an hour)	best (best)	bad (bad)	best (best)
bala (without)	allah (God)	se3a (zero)	ahla (best)
khara (shit)	ahla (best)	bala (without)	ktir (much)
bad (bad)	7elo (nice)	la (no)	allah (God)
3ayb (disgrace)	perfect (perfect)	mara (once)	tayeb (delicious)

4.3.1.2 ***Second phase (hyperparameters tuning settings)***

The second phase in the second experiment, another two LR models with BoW, TF*IDF, and n-grams features enhanced through hypermeters tuning, trained and tested on the same split made on the first experiment to maintain parallel results. To achieve the best of LR performance,

we implemented *gridsearchcsv*[12] from *scikit-learn* on the LR model by creating a pipeline that facilitate the LR's hyperparameters optimization. We have given choices for each parameter that undertaken in both models, we specify them in the following Table (24) for BoW and TF*IDF:

***Table 24**- Pipeline Hyperparameters tuning architecture*

Parameter\architecture	**BoW LR pipeline**	**TF*IDF LR pipeline**
multi-level n-gram	1) unigram (1,1) 2) unigram and bigram (1,2) 3) unigram, bigram, and trigram (1,3)	
inverse of regularization strength	100000, 10000, 1000, 100, 10, 1, 0.1, 0.01, 0.001, 0.0001	
cross validation	2	
scoring	f1-score	

The pipeline architecture in Table (24) specifies n-gram range is in unigram (1,1), unigram and bigram (1,2), and unigram, bigram, and trigram (1,3), [10**-i for i in range(-5, 5)] regularize strength, 2 cross validation folds with a conditioned evaluation on f1-score performance in order to output data-fitted model most effectively when it reaches to the highest f1-score, and the remaining parameters stayed default as the have been specified earlier in the first phase models.

The running of the pipeline experiments for both LRs models result in the following choices for both pipelines:

***Table 25**- Pipeline Hyperparameters tuning Best Model Architecture*

Parameter\architecture	**BoW LR**	**TF*IDF LR**
multi-level n-gram	unigram (1,1)	unigram (1,1)
Inverse of regularization strength	10	10

[12] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

scoring	0.896	0.909

4.3.2 **Lexicon-based**

To build the lexicon-based classifier, we need first to create the main components of SCLSAS, which are the dictionary, grammar rules, and the semantic map categories. We will discuss each in the following subsections.

4.3.2.1 ***Dictionary***

The dictionary is manually built, collected, and annotated through deep analysis on 25% of the training set that was outputted in the 4.1 **Data Preprocessing** step. The dictionary contains a set of terms that are in two sentiment classes: positive and negative. For both classes, we provided each term to a specific subdomain in the assigned class by means of professional sentiment classification. Each target term specified in its domain with a linguistics marker that happens to come before or after the occurrence of it. The total number of both markers and target terms in both negative and positive classes are presented in the following Table (26).

Table 26*- Dictionary Categories For Terms in Postive & Negative Classes*

Category\Class	**#Positive Class Terms Markers**	**#Positive Class Target Terms**	**#Negative Class Terms Markers**	**#Negative Class Target Terms**
Delivery	1	1	16	19
Costumer service	22	21	24	22
Price	14	18	9	10
Recommendation & Suggestion	62	66	46	47
Administration	99	126	65	86
Service	33	41	38	42

Product	106	131	64	84
Market	36	40	34	32
Ambiance	61	77	6	6
Overall	45	88	37	51

4.3.2.2 ***Grammar rules (linguistics rules)***

After building the terminological dictionary of positive and negative classes, grammar rules are specified for the classifier to follow in accordance to sentiment classification of Arabizi text reviews. In each sentiment class, there are 11 categories for term entry that was specified above in Table (26). And, each category there is a term marker and a target one. Accordingly, the grammar is derived from this by specifying markers of a category followed by possible target terms. The rules are written for each one of the categories in both positive and negative classes. An example of delivery category in the positive class would be specifying the marker with target sentiment text that is extracted from SLCSAS as follows:

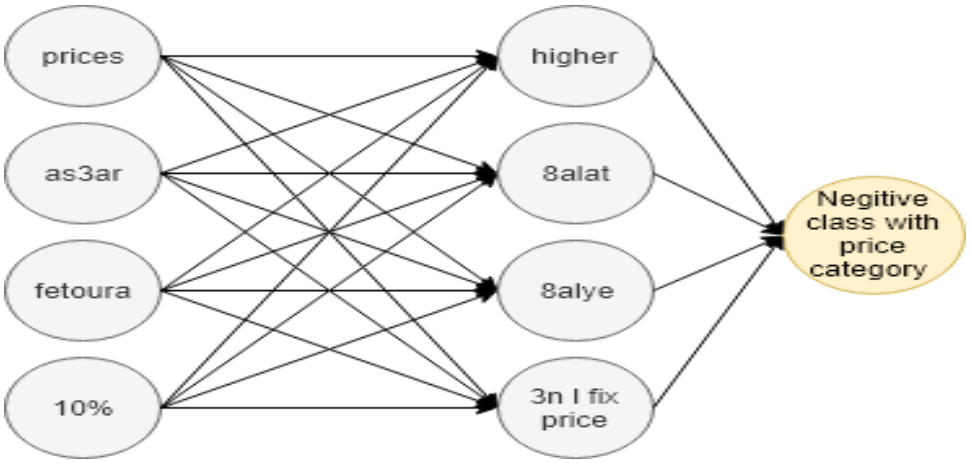
Figure 19*- A Simple Example of Grammar Rule in the Delivery Category*

```
::pos_marker_delvry = (delivery)

::pos_txt_delvry = (sari3)

::pos_marker_delvry > ::pos_txt_delvry -> pos_delivery
```

From the above example two variable were constructed to recognize both positive delivery marker and the sentiment target text that should be followed with. The third line remarks the occurrence of the term marker delivery and the following of the target sentiment text sari3/fast tagging as positive class with category of delivery.

However, grammar rules are much more complicated than what is the previous example. Markers are listed in SLCSAS variables without a chain to specific sentiment target texts but to multiple terms in target.

***Figure 20**- Complex Example of Grammar Rule in the Price Category of the Negative Class*

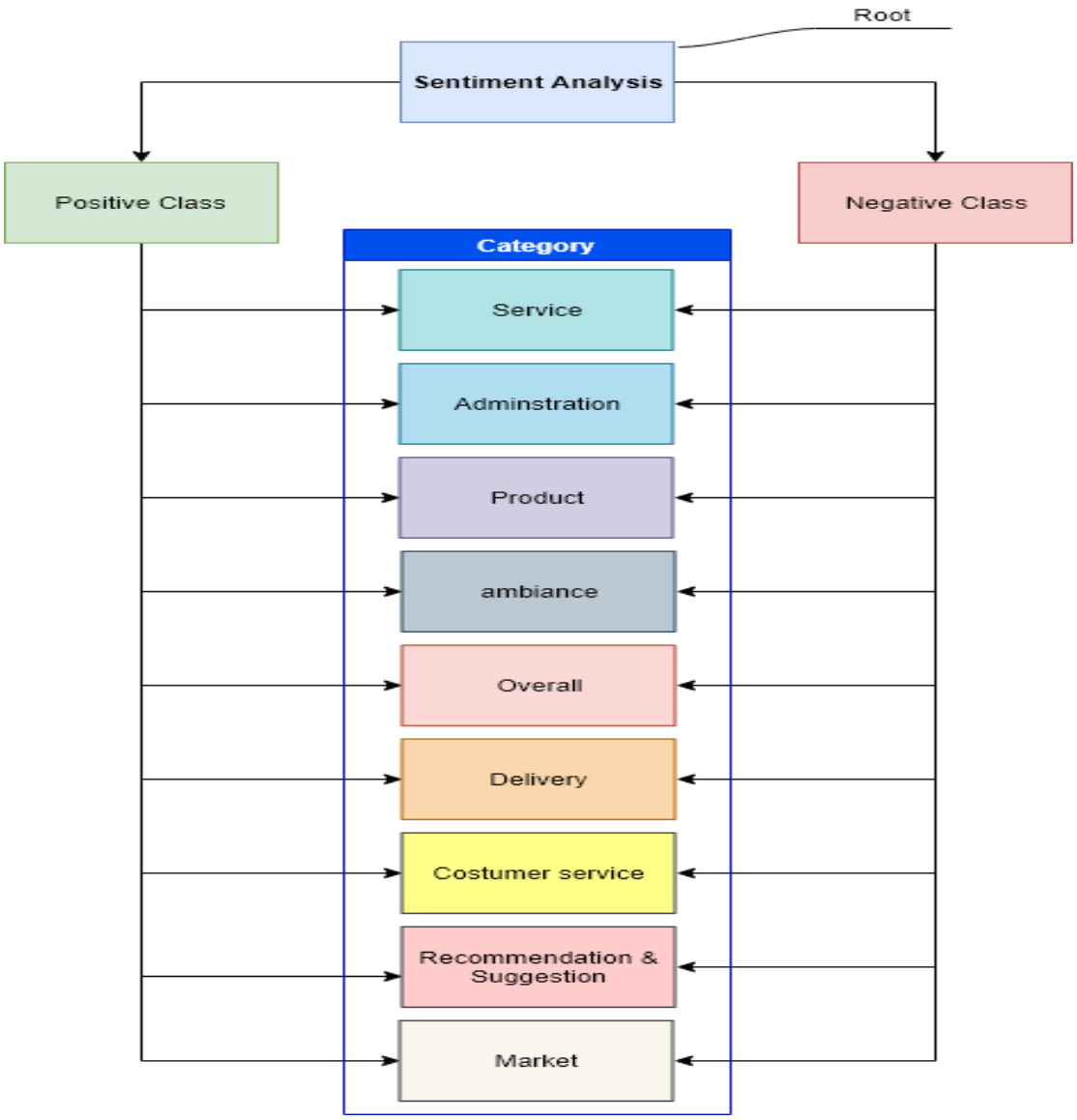


From the above Figure (19), it shows a complex relationship between markers of prices, as3ar/prices, fetoura/check, and 10% to sentiment texts in target, including higher, 8alat/false, 8alye/expensive, and 3n l fix price/than the fix price. In practice, if one term of the list has occurred with a target sentiment text, the predication would be negative class with price category.

4.3.3.3 ***Semantic map***

The semantic map specifies the root of SLCSAS classifier, which is SA with the sentiment classes of positive and negative, and the categories that are specified for each class in the dictionary.

***Figure 21**- Semantic Map of both positive and negative classes with corresponding category*



4.4 **Results and Evaluation**

Both lexicon-based and ML models experimented and evaluated on the datasets over a single laptop that has the following specification:

***Table 27**- Computer Specification*

OS Name	Microsoft Windows 10 Home
System Type	x64-based PC

Processor	Intel(R) Core(TM) i7-6700HQ CPU @ 2.60GHz, 2601 Mhz, 4 Core(s), 8 Logical Processor(s)
VGA1	Intel HD Graphics 530
VGA2	GeForce GTX 960M 4GB
RAM	12 GB

For results' evaluations, we compare the performance of both models approaches on precision, recall, f1-score, confusion matrix, and Receive Operating Characteristics (ROC) measures. These measures assess both classifiers in terms of binary classification problem (positive or negative sentiment classes).

1) **Precision and Recall**: precision and recall are useful measures in IR tasks where precision is the measure of result relevancy, whereas recall measures how many relevant results have been returned. Precision (P) calculated by dividing true positives (T_P) over the number of true positives (T_P) plus the number of false positives (F_P).

$$P = \frac{T_P}{T_P + F_P} \tag{17}$$

Recall (R) calculated by dividing number of true positives (T_P) over the number of true positives (T_P) plus the number of false negatives (F_N).

$$R = \frac{T_P}{T_P + F_N} \tag{18}$$

2) **F1-score**: Both precision and recall are related to $F_1 - \text{score}$, which is the weighted average of $P \; and \; R$.

$$F1 - Score = 2\frac{P \times R}{P + R} \tag{19}$$

3) **Confusion matrix**: it is useful for describing the performance of classification tasks on test data with known true values. The matrix uses four main rates of true positives (T_P), true negatives (T_N), false positives (F_P), and false negatives (F_N), as specified in the following table:

***Table 28**- Confusion Matrix in Binary Classification Task*

Actual\Predicated	**Predicted: Negative**	**Predicted: Positive**
Actual: Negative	T_N	F_P
Actual: Positive	F_N	T_P

4) **Receive Operating Characteristics (ROC) curve**: It is a metric for binary classification problem, which considers all possible thresholds. Various thresholds' values result in different true positive (T_P)/false positive (F_P) rates. In addition, The area under the curve (AUC) represents how skillful the model is by computing F_P rate (FPR) and T_P rate (TPR), where FPR is computed as the number of F_P divided by the sum of the number of F_P and the number of T_N, and (TPR)/sensitivity is calculated as the number of T_P divided by the number of T_P and the number of F_N (Brownlee, 2018). Both variables presented in the following equation descriptors:

$$Sensitivity/TPR/P(T_P) = \frac{T_P}{T_P+F_N} \quad (20)$$

$$FPR/P(F_P) = \frac{F_P}{F_P+T_N}, \quad (21)$$

where FPR is referred to as the inverted specificity ($1 - specificity$) that it is the total number of T_P divided by the sum of T_N and F_P.

$$Specificity = \frac{T_N}{T_N+F_P} \quad (22)$$

V. Research Result

This section represents the experiments and results of conducted experiments for the dataset on public and private services' reviews in the Lebanese Arabizi. Since we have more than one ML models and rule-based sentiment classifiers, experiments for each classifier were carried out.

5.1 **Machine Learning**

ML experimentation is into two primary phases in which the first phase is for training LR classifier with default settings as specified earlier in 4.3.1.1 **First phase (Default settings)**, and the second phase is for training a LR classifier with hypermeters tuning settings as specified in detail in 4.3.1.2 **Second phase (hyperparameters tuning settings)**.

5.1.1 **First phase (Default settings)**

The results of both models at the first phase are presented in Tables (29) and (30). We observe that the precision, recall, f1-score obtained by BoW feature-based classifier is 0.84% 0.62% 0.71% for negative class and 0.83% 0.94% 0.88% for positive class respectively. On the other hand, we observe that the precision, recall, and f1-score obtained by TF*IDF feature-based classifier is 0.82% 0.63% 0.71% for negative class and 0.83% 0.93% 0.88% for positive class respectively.

***Table 29**- Performance of BoW LR Model with Default Settings*

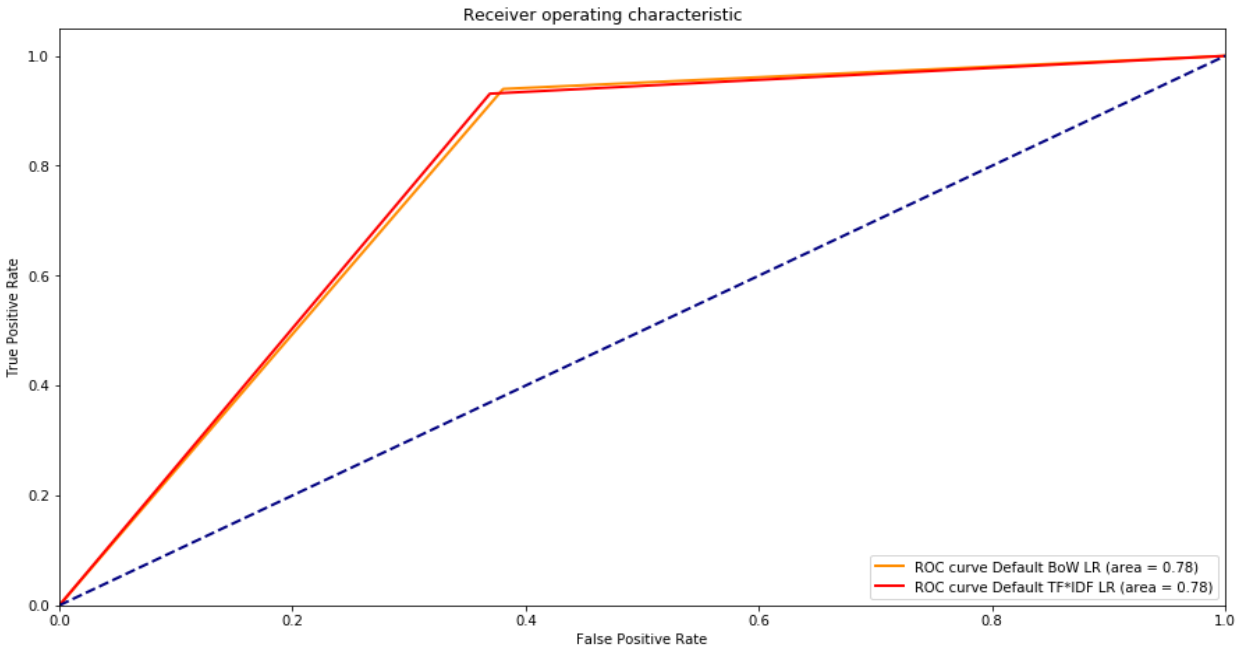
Class\Measure	precision	recall	f1-score	support
0	0.84	0.62	0.71	176
1	0.83	0.94	0.88	349

***Table 30**- Performance of TF*IDF LR Model with Default Settings*

Class\Measure	precision	recall	f1-score	support
0	0.82	0.63	0.71	176
1	0.83	0.93	0.88	349

In the following figure, we present the achieved results of the default LR classifiers without hypermeter tuning in AUC representation.

***Figure 22**- The Receiver Operating Characteristics curves of BoW and TF*IDF LR Models with Default Settings*



We can observe from the above figure a slight shift from LR model trained with TF*IDF feature and n-grams to the one trained with BoW feature. The difference in the AUC of about 3 points shift has a significance results also reflected on the rates of False Positive (FP) and (True Positive). The BoW LR model outperforms TF*IDF one, as they both reflect competitive scores on the test data as it is presented in the following tables.

***Table 31**- Confusion Matrix of BoW LR Model with Default Settings*

Actual\Predicated	**Predicted: Negative**	**Predicted: Positive**
Actual: Negative	109	67
Actual: Positive	21	328

***Table 32**- Confusion Matrix of TF*IDF LR Model with Default Settings*

Actual\Predicated	**Predicted: Negative**	**Predicted: Positive**

Actual: Negative	111	65
Actual: Positive	24	325

5.1.2 **Second phase (hyperparameters tuning settings)**

We adopted the same evaluation measures used in the first experiment. The outcomes of both models are presented in Tables (33) and (34). We observe that the precision, recall, and f1-score gotten by optimized BoW feature-based classifier is 0.86% 0.69% 0.76% for negative class and 0.86% 0.94% 0.90% for the positive examples. On the other hand, the TF*IDF feature-based LR achieved precision, recall, and f1-score of 0.80% 0.75% 0.77% for negative class and 0.88% 0.90% 0.89% for negative sentiment respectively.

***Table 33**- Performance of BoW LR Model with Hypermeters Tuning Settings*

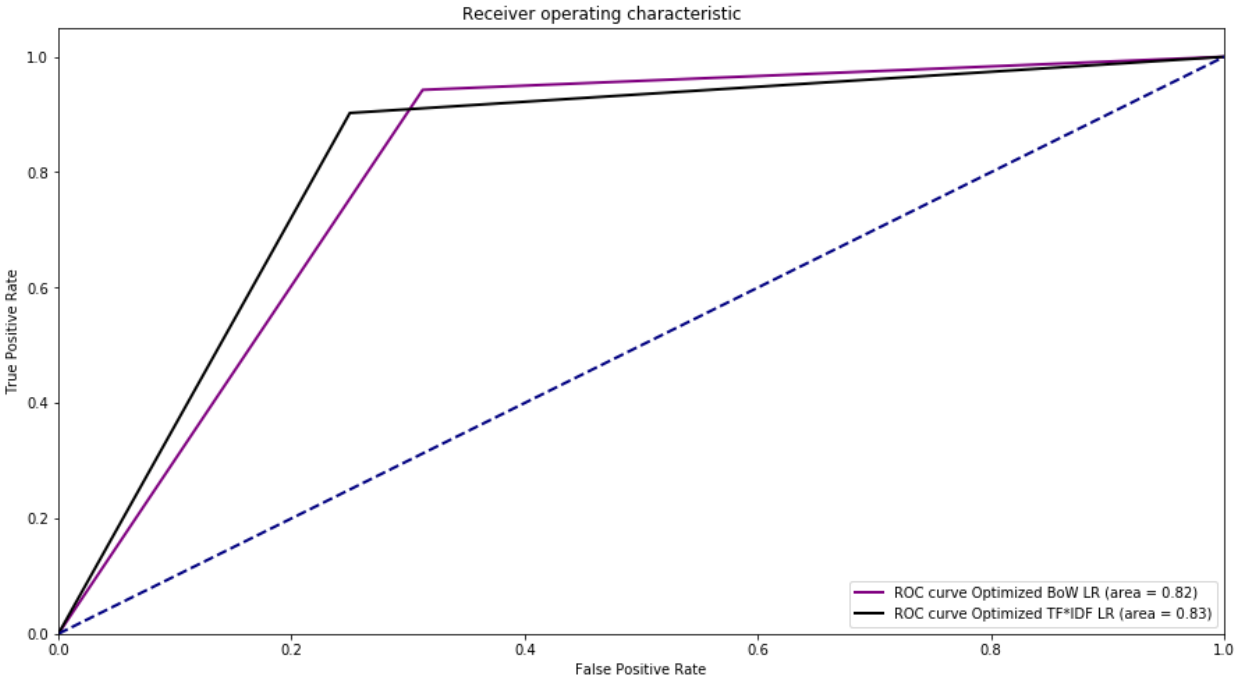
Class\Measure	**precision**	**recall**	**f1-score**	**support**
0	0.86	0.69	0.76	176
1	0.86	0.94	0.90	349

***Table 34**- Performance of TF*IDF LR Model with Hypermeters Tuning Settings*

Class\Measure	**precision**	**recall**	**f1-score**	**support**
0	0.80	0.75	0.77	176
1	0.88	0.90	0.89	349

In the following figure, we present the achieved results of the hypermeters tuned LR classifiers with BoW and TF*IDF in AUC representation.

***Figure 23**- The Receiver Operating Characteristics curves of BoW and TF*IDF LR Models with Hypermeters Tuning Settings*



We can observe from the AUC plot of the two models a distinguished significance from LR model trained with TF*IDF feature and n-grams to the one trained with BoW feature. This puts LR with TF*IDF feature on the top that reflects competitive scores on the test data as with rates of False Positive (FP) and True Positive (TP), presented in the following tables.

***Table 35**- Confusion Matrix of BoW LR Model with Hypermeters Tuning Settings*

Actual\Predicated	**Predicted: Negative**	**Predicted: Positive**
Actual: Negative	121	55
Actual: Positive	20	329

***Table 36**- Confusion Matrix of TF*IDF LR Model with Hypermeters Tuning Settings*

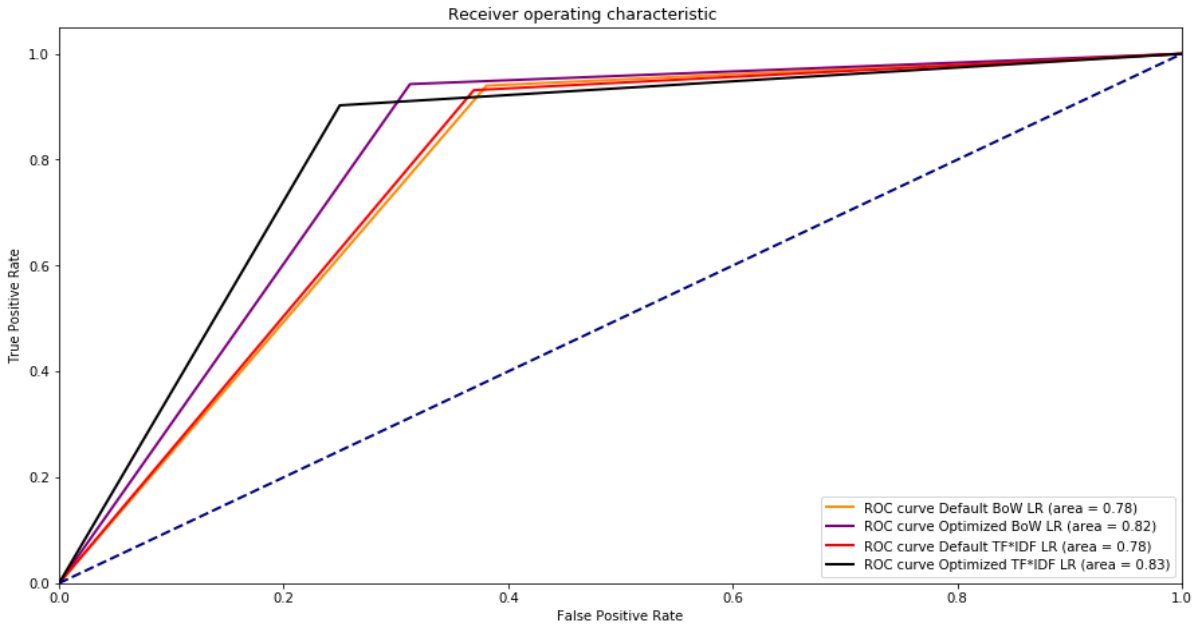
Actual\Predicated	**Predicted: Negative**	**Predicted: Positive**

Actual: Negative	132	44
Actual: Positive	34	315

5.1.3 **Experiment Summary**

In the first experiment, two ML models were applied to our dataset that consist of reviews from Google, Facebook, and Zomato. As it was discussed in section, the data itself should be represented in vector space. In purpose to find the optimal data representation, experiments were conducted with BoW of unigram, bigram, and trigram word levels dictionary, and TF*IDF of unigram, bigram, and trigram word levels dictionary, too. Both BoW and TF*IDF models showed a similar result with precision, recall, f1-score, 0.84% 0.62% 0.71% for negative class and 0.83% 0.94% 0.88% for positive class respectively by LR with BoW feature. On the other hand, we observe that the precision, recall, and f1-score obtained by TF*IDF feature-based LR classifier is 0.82% 0.63% 0.71% for negative class and 0.83% 0.93% 0.88% for positive class respectively on the test data. In the second experiment, two other LR models have been hypermeters tuned to reach to maximize sentiment classification performance on the same dataset. Experiments showed, choosing inverse of regularization strength to be equal to 10, and TF*IDF with unigram word level n-gram outperforms LR hyperparameters tuned model with 10 inverse of regularization strength and unigram BoW. The TF*IDF feature-based LR achieved precision, recall, and f1-score of 0.80% 0.75% 0.77% for negative class and 0.88% 0.90% 0.89% for negative sentiment respectively, whereas optimized BoW feature-based classifier is 0.86% 0.69% 0.76% for negative class and 0.86% 0.94% 0.90% for the positive examples respectively.

***Figure 24**- Summary: The Receiver Operating Characteristics curves of First- and Second-ML Phases of LR*



5.2 **Lexicon-based**

SLCSAS, implemented on the test set made in ML, remarks the following occurrences 11 semantic map categories that detailed in the following Table (37) with review representation of positive class categories in 219 reviews while negative categories occurred in 101.

***Table 37**- Semantic map Categories found in test data*

Reviews Representation	N = 219	N = 101
Category\Class	**Positive Class**	**Negative Class**
Delivery	0	5
Costumer service	3	5
Price	6	3
Recommendation & Suggestion	12	14
Administration	60	11

Service	46	28
Product	136	41
Market	46	20
Ambiance	66	1
Overall	52	29

We adopted the same evaluation measures used in the ML implementation. The outcomes of conducting the rule-based experiment is presented in Table (38). We observe that the classifier was able to recognize 274 text reviews (52.19%) out of 525 in total. Moreover, the precision, recall, and f1-score achieved by SLCSAS classifier is 0.616% 0.803% 0.697% for negative class and 0.841% 0.93% 0.883% for the positive examples respectively.

***Table 38**- Performance of SLCSAS Classifier on 274 text reviews in total*

Class\Measure	**precision**	**recall**	**f1-score**	**retrieved**	**support**
0	0.616	0.803	0.697	86	176
1	0.841	0.93	0.883	188	349

Moreover, the 274 retrieved text reviews have been evaluated manually through comparison between the actual sentiment and the predicated categories class. The following part specifies 4 examples in $T_P, F_P, T_N, and\ F_N$ noted in the SLCSAS classifier's results with the overall confusion matrix evaluation regarded in Table (39).

***Figure 25**- SLCSAS F_P Example*

***Figure 26**- SLCSAS F_N Eample*

***Figure 27**- SLCSAS T_p Eample*

***Figure 28**- SLCSAS T_N Eample*

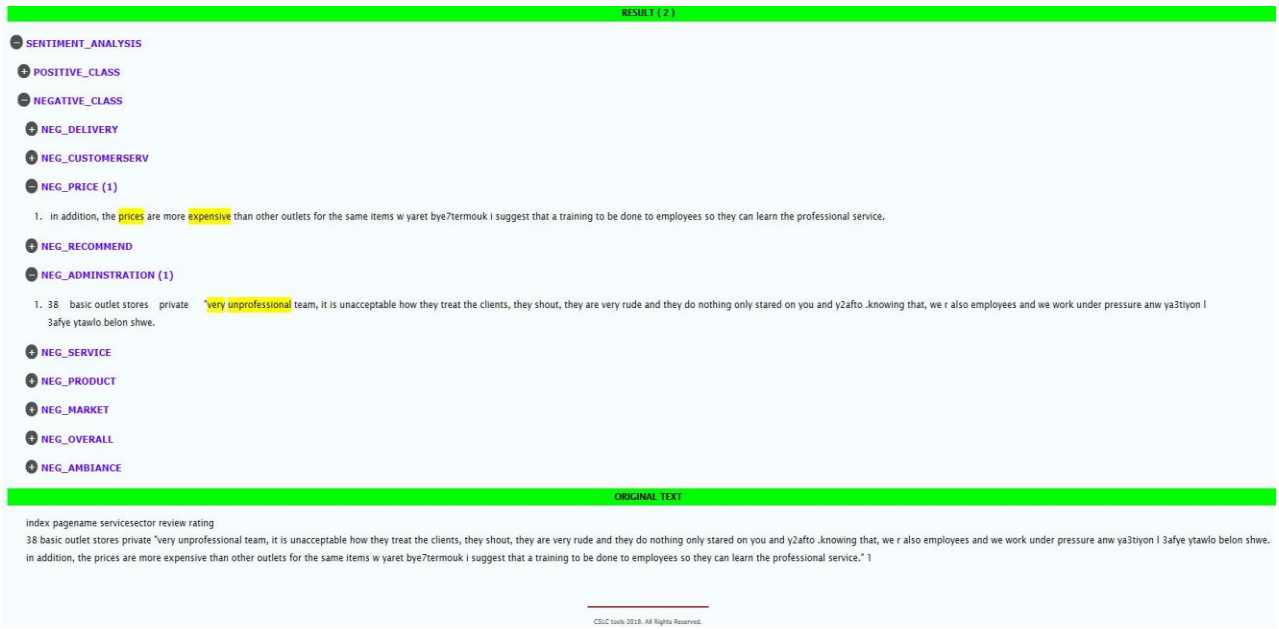


***Table 39**- Confusion Matrix of the Results of SLCSAS*

Actual\Predicated	Predicted: Negative	Predicted: Positive
Actual: Negative	53	33
Actual: Positive	13	175

5.2.1 **Experiment Summary**

In the linguistic-rule classifier, SLCSAS was used with a combination of lexicon dictionary of positive and negative terms markers and target sentiment ones represented in grammatical rules for the classification tagging. The semantic map tagged with 11 categories for each of the sentiment classes positive and negative ones. On test data experiment, SLCSAS was able to capture 274 reviews out of 525. the achieved results in terms of precision, recall, and f1-score by SLCSAS classifier are 0.616% 0.803% 0.697% for negative class and 0.841% 0.93% 0.883% for the positive examples respectively. The classifier achieved in end performance in terms of confusion matrix measure the following: 175 T_P, 33F_P, 53 T_N, *and* 13 F_N.

Positive categories achieved a total number of 0 delivery, 3 customer service, 6 price, 12 recommendation & suggestion, 60 administration, 46 service, 136 product, 46 market, 66 ambiance, and 52 overall in 219 reviews occurrence. On the other hand, negative categories achieved a total number of 5 delivery, 5 customer service, 3 price, 14 recommendation & suggestion, 11 administration, 28 service, 41 product, 20 market, 1 ambiance, and 29 overall in 101 reviews occurrence.

5.3 **Discussion**

This section deals with discussion over the results and findings that described in the previous section of experiment. First, we compare the results of different classifiers applied for SA of services' reviews obtained from Google, Facebook, and Zomato platforms. Second, the discussion of the impact of different features are presented. Finally, we discuss the best obtained results obtained by optimized TF*IDF LR Model.

***Table 40**- Performance Comparison Between ML and Lexicon-based classifiers*

Model	Class\Measure	precision	recall	f1-score	retrieved	support
Unoptimized BoW LR	0	0.84	0.62	0.71		
	1	0.83	0.94	0.88		
Unoptimized TF*IDF LR	0	0.82	0.63	0.71		
	1	0.83	0.93	0.88	176	
Optimized BoW LR	0	0.86	0.69	0.76	349	
	1	0.86	0.94	0.90		
Optimized TF*IDF LR	0	0.80	0.75	0.77		
	1	0.88	0.90	0.89		
SLCSAS	0	0.616	0.803	0.697	86	176
	1	0.841	0.93	0.883	188	349

Based on Table (38), both the unoptimized LR models of BoW manifest similar results with a slight divergence in scores. The f1-score of both is similar, which the tread-off between precison and recall. On the other hand, both the optimized LR models remark a slight difference in the achieved f1-score with a higher performance than the ones achieved in the unoptimized models. The results of BoW LR model in negative class predication jumped from 0.71 to 0.76 with 0.05 f1-score difference. In addition, the positive class predication jumped from 0.88 to 0.90 with 0.02 f1-score difference. Besides, the results of TF*IDF LR model in negative class predication shifted 0.71 to 0.77 with 0.06 f1-score difference, whereas the positive class predication remarked a 0.01 difference from 0.88 to 0.89 f1-score.

On the contrary, SLCSAS achieved a good f1-socre of 0.697 for negative class and 0.883 for positive one. For positive sentiment class, SLCSAS is comparable to the unoptimized LR models with BoW and TF*IDF, whereas for the negative class it is comparable to the unoptimized

BoW model but with flipped values of precision and recall. The BoW LR model achieved 0.84 precision and 0.62 recall for negative class, while it is 0.616 precision and 0.803 recall in the SLCSAS. It is a note that both recalls of positive and negative classes are high, which are quite remarkable to ML models. However, predication coverage still be lacked in SLCSAS linguistic-based classifier unlike what is in the ML models that cover all the test data. This may go back to the full training of ML models on the training set, while it is 25% hand-crafted training on the training set left the SLCSAS classifier downward to ML ones.

Secondly, ML results show that data representation features have achieved similar performance with a slight increase and decrease in scores. The Unoptimized models have showed same results in terms of f1-score, while it has increased in the optimized TF*IDF LR model from the BoW one with $0.01\ (0.77 - 0.76)$ and $0.01\ (0.90 - 0.89)$ f1-scores for negative and positive classes respectively.

VI. Conclusion

This research work represents comparative SA approaches applied to services' reviews corpus gathered from Google, Facebook, and Zomato websites, and aims to compare both standard and optimized ML approach techniques with lexicon-based approach.

To answer the research questions, we conducted five experiments of multiple ML models paired with different features and optimization techniques besides lexicon-based classifier, experiments for each model were carried out. First, the ML approach has achieved better results than those achieved by lexicon-based classifier. Secondly, we have found that the best model and classifier that fits binary SA task is optimized LR model trained through TF*IDF feature with word level unigram and 10 inverse of regularization strength. Also, we have found that ratings' encodings to ones and zeroes for positive and negative classes respectively, data percentage split, lower casing the text reviews, and finally removing of reviewers' related information in the corpus are possible candidates for data preprocessing in the Arabizi NLP in binary SA task. Finally, we have found that the best ML feature is both BoW and TF*IDF with world level n-grams (unigrams, bigrams, and trigrams) in the case of standard LR implementation, whereas it is TF*IDF feature model in the case of optimized LR implementation.

To find out about the hypotheses that we set earlier before doing the research experiments, we found that LR ML model is applicable to classify Arabizi texts in binary classification problems in our case it is positive or negative classes. Moreover, the second hypothesis held true because we were able to find the target text reviews to implements the experiments on. Next, the built corpus is hugely classified in private service sector, whereas there is little found in the public one, which result in undigitized governmental society. And, we have found an opposite result to the fourth hypothesis. The experiments show a tremendous difference in the achieved results in ML and lexicon-based classifiers. Finally, we have found that the ML approach remarks a better result than rule-based approach, which in/validates the last hypothesis.

6.1 **Future Work**

This thesis opens up further research opportunities on SA task for Arabizi ALP, using both ML algorithms and rule-based ones paired with various data features other than the ones taken in this study. Moreover, it is worth to expand the corpus to include much more reviews in a wide scope of regions other than Lebanon to have a much more generalizable performance and data scalability. It is also considerable to expand the research on Arabizi for further tasks and subjects as for example part-of-speech tagging and sentence categorization. Also, using DL architectures are to be considered for future work particularly by using Long Short-Term Memory (LSTM) and other data representation techniques as word embedding, for example, word2vec (Mikolov et al., 2013).

References

Abdualla, N., Ahmad, N., Shehab, M., & Al-Ayyoub, M. (2013). Arabic Sentiment Analysis: Lexicon-based and Corpus-Based. 2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), Amman, Jordan, pp. 1-6.

Aboelezz, M. (2012). We are young. We are trendy. Buy our product! The use of Latinized Arabic in edited printed press in Egypt. United Academics Journal of Social Sciences, 2, 48-72. Retrieved September 10, 2018, from http://www.lancs.ac.uk/pg/aboelezz/docs/WeAreYoungTrendy.pdf

Agarwal, B., Mittal, N., Bansal, P., & Garg, S. (2015). Sentiment Analysis Using Common-Sense and Context Information. Computational Intelligence and Neuroscience., vol. 2015, pp. 1–9.

Agrawal, R., Rajagopalan, S., Srikant, R., & Xu, Y. (2003). Mining newsgroups using network arising from social behavior. Twelfth international World Wide Web Conference.

Al Omari, M. & Al-Hajj, M. (2019). Classifiers for Arabic NLP: Survey. Int. J. of Computational Complexity and Intelligent Algorithms (IJCCIA). Doi: 10.1504/IJCCIA.2018.10019805

Al Omari, M., Al-Hajj, M., Hammami, N., & Sabra, A. (2019). Sentiment Classifier: Logistic Regression for Arabic Services' Reviews in Lebanon. 2019 International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia, 2019, pp. 1-5. Doi: 10.1109/ICCISci.2019.8716394

Al Omari, M., Al-Hajj, M., Sabra., A. (2019). Hybrid CNNs-LSTM Deep Analyzer for Arabic Opinion Mining. The International Conference on Arabic Language Processing (ICALP), Nancy, France.

Al-Hajj, M. (2018). SLC Semantic Analysis System (SLCSAS). Retrieved February 21, 2019, from https://github.com/moustafaalhajj/SLCSAS

Al-Smadi, M., Al-Ayyoub, M., Jararweh, Y., Qawasmeh, O., Benkhelifa, E. and Talafha, B. (2018). Enhancing Aspect-Based Sentiment Analysis of Arabic Hotels' Reviews Using Morphological, Syntactic and Semantic Features. Journal of Information Processing & Management. DOI: 10.1016.

Arthur, S. (1959). Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development. 3 (3), pp. 210–229. CiteSeerX 10.1.1.368.2254. Doi:10.1147/rd.33.0210.

Bakshi, K. (2012). Considerations for Big Data: Architecture and Approaches. In: Proceedings of the IEEE Aerospace Conference, pp. 1–7.

Bansal, S. (2018). Turing Test in Artificial Intelligence. Retrieved February 21, 2019, from https://www.geeksforgeeks.org/turing-test-artificial-intelligence

Basis Technology. (2012). The burgeoning challenge of deciphering Arabic chat, pp. 1-9.

Bastien, L. (2018). Data Mining : Qu'est ce que l'exploration de données ? Retrieved October 17, 2018, from https://www.lebigdata.fr/data-mining-definition-exemples

Bhatti, Z., Waqas, A., Ismaili, I., Hakro, D., & Soomro, W. (2014). Phonetic based SoundEx & ShapeEx algorithm for Sindhi Spell Checker System. American Eurasian Network for Scientific Information: Advances in Environmental Biology (AENSI-AEB), 8 (4), pp. 1174-1155.

Bianchi, R. M. (2012). 3arabizi - When Local Arabic Meets Global English On The Internet. Acta Linguistica Asiatica, 2(1), pp. 89-100.

Big Data Analytics. (n.d.). Retrieved October 1, 2018, from https://www.ibm.com/analytics/hadoop/big-data-analytics

Bishop, M. (2006), Pattern Recognition and Machine Learning, Springer, ISBN 978-0-387-31073-2

Book Reviews. (n.d.). Retrieved October 1, 2018, from https://writingcenter.unc.edu/tips-and-tools/book-reviews/

Brownlee, J. (2016). Supervised and Unsupervised Machine Learning Algorithms. Retrieved October 31, 2018, from https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms

Brownlee, J. (2018). How and When to Use ROC Curves and Precision-Recall Curves for Classification in Python. Retrieved March 31, 2019, from https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/

Cotterell, R., Renduchintala, A., Saphra, N., & Callison-Burch, C. (2014). An Algerian Arabic-French Code-Switched Corpus.

Cramer, J. (2002). The Origins of Logistic Regression. inbergen Institute Working Paper No. 2002-119/4, pp. 1-16.

Crystal, D. (2001). Language and the Internet. Cambridge: Cambridge University Press.

Duwairi, R., Alfaqeh, M., Wardat, M., & Alrabadi, A. (2016). Sentiment Analysis for Arabizi Text. Proceedings of the 7th International Conference on Information and Communication Systems (ICICS), Amman, Jordan, pp. 127-132.

El-Beltagy, S., & Ali, A. (2013). Open Issues in the Sentiment Analysis of Arabic Social Media: A Case Study. Proceedings of the 9th International Conference on Innovations in Information Technology (IIT), Abu Dhabi, United Arab Emirates, pp. 215-220.

Elegendy, N., & Elragal, A. (2014). Big Data Analytics: A Literature Review Paper. P. Perner (Ed.): ICDM 2014, LNAI 8557, pp. 214–227.

Elgendy, N., & Elragal, A. (2014). Big Data Analytics: A Literature Review Paper. In: Perner P. (eds) Advances in Data Mining. Applications and Theoretical Aspects. ICDM 2014. Lecture Notes in Computer Science, vol 8557, pp. 214-227. Springer, Cham. DOI: https://doi.org/10.1007/978-3-319-08976-8_16

Farra, N., Challita, E., Assi, R., & Hajj, H. (2010). Sentence-Level and Document-Level Sentiment Mining for Arabic Texts. *IEEE International Conference on Data Mining Workshops*, Sydney, NSW, 2010, pp. 1114-1119. Doi: 10.1109/ICDMW.2010.95.

Feldman, R. (2013). Techniques and Applications for Sentiment Analysis. Commun. ACM 56(4), pp. 82–89.

Gantz, J., & Reinsel, D. (2012). Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East Executive Summary: A Universe of Opportunities and Challenges, Idc, vol. 2007, no. December 2012, pp. 1–16.

Guestrin, C., & Fox, E. (2016a). Logistic regression model. Retrieved February 21, 2019, from https://www.coursera.org/lecture/ml-classification/logistic-regression-model-OJQXu

Guestrin, C., & Fox, E. (2016b). Predicting class probabilities with (generalized) linear models. Retrieved February 21, 2019, from https://www.coursera.org/lecture/ml-classification/predicting-class-probabilities-with-generalized-linear-models-OV5Kt

Guestrin, C., & Fox, E. (2016c). Example of computing derivative for logistic regression. Retrieved March 25, 2019, from https://www.coursera.org/learn/ml-classification/lecture/UEmJg/example-of-computing-derivative-for-logistic-regression

Heikal, M, Torki, M., and El-Makky, N. (2018). Sentiment Analysis of Arabic Tweets using Deep Learning. Procedia Computer Science 142, pp. 114-122. DOI: 10.1016/j.procs.2018.10.466

Hu, M., & Liu, B. (2004). A list of positive and negative opinion words or sentiment words for English. Retrieved October 26, 2018: from http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon

Hutchins, J. (2004) The Georgetown-IBM Experiment Demonstrated in January 1954. In: Frederking R.E., Taylor K.B. (eds) Machine Translation: From Real Users to Research. AMTA 2004. Lecture Notes in Computer Science, vol 3265. Springer, Berlin, Heidelberg, pp. 102-114.

Hutchins, J., & Hays, G. (2015). 11 ALPAC : The (In) Famous Report. Retrieved February 21, 2019, from https://www.semanticscholar.org/paper/11-ALPAC-%3A-The-(-In-)-Famous-Report-Hutchins-Hays/7c0a06f3d5cddd2bf9796b92d1a03d366aa9351b

Johnson, M. (2009). How the statistical revolution changes (computational) linguistics. In Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous? (ILCL '09). Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 3-11.

Khayyat, M. (1999). The very resistable rise of Arabinglizi. Retrieved October 22, 2018, from http://www.dailystar.com.lb/Culture/Art/1999/Jun-15/98190-the-very-resistable-rise-of-arabinglizi.ashx

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2017). Natural Language Processing: State of The Art, Current Trends and Challenges. CoRR, abs/1708.05148.

Kohavi, R., & Provost, F. (1998). Glossary of terms. Special Issue on Applications of Machine Learning and the Knowledge Discovery Process. Machine Learning, 30, pp. 271–274. Retrieved October 30, 2018, from http://robotics.stanford.edu/~ronnyk/glossary.html

Larousse, É. (n.d.). Définitions : Classifier - Dictionnaire de français Larousse. Retrieved October 1, 2018, from https://www.larousse.fr/dictionnaires/francais/classifier/16416

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature Publishing Group, a division of Macmillan Publishers Limited, pp. 436–444. DOI: https://doi.org/10.1038/nature14539

Manning, C. D., Raghavan, P., & Schütze, H. (2009). An Introduction to Information Retrieval (Online ed.). Retrieved October 27, 2018, from https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf

Middle East Internet Statistics, Population, Facebook and Telecommunications Reports. (2018). Retrieved September 25, 2018, from https://www.internetworldstats.com/stats5.htm

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems. arXiv:1310.4546.

Mitchell, T. (1997). Machine Learning. McGraw Hill. ISBN 978-0-07-042807-2.

Morsy S., & Rafea. A. (2012). Improving Document-Level Sentiment Classification Using Contextual Valence Shifters. Natural Language Processing and Information Systems Lecture Notes in Computer Science Volume 7337, 2012, pp. 253-258.

Mouthami, K., Devi, K. N., & Bhaskaran, V. M. (2013). Sentiment Analysis and Classification Based on Textual Reviews. In: International Conference on Information Communication and Embedded Systems (ICICES), pp. 271–276.

Muhammed, R., Farrag, M., Elshamly, N., and Abdel-Ghaffar, N. (2011). Arabizi or Romanization: The dilemma of writing Arabic texts. Jīl Jadīd Conference, University of Texas at Austin, February 18-19, 2011, The American University in Cairo (VIA SKYPE). Retrieved October 20, 2018, from https://ecitydoc.com/download/summary-of-arabizi-or-romanization-the-dilemma-of-writing-arabic_pdf

Muysken, P. (2000). Bilingual speech: A typology of code-mixing. Cambridge University Press.

Noam, C. (1965). Aspects of the Theory of Syntax. MIT Press. ISBN 0-262-53007-4.

Oudah, M., & Shaalan, K. (2012). A Pipeline Arabic Named Entity Recognition Using a Hybrid Approach. Proceedings of COLING 2012: Technical Papers, Mumbai, pp. 2159-2176.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, Vol. 2, Nos. 1–2, pp. 1-135. DOI:10.1561/1500000011.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.79–86.

Petasis, G., Vichot, F., Wolinski, F., Paliouras, G., Karkaletsis, V., & Spyropoulos, C. D. (2001). Using Machine Learning to Maintain Rule-based Named-Entity Recognition and Classification Systems. Proceeding Conference of Association for Computational Linguistics, pp. 426-433.

Salem, F. (2017). The Arab Social Media Report 2017: Social Media and the Internet of Things: Towards Data-Driven Policymaking in the Arab World (Vol. 7). Dubai: MBR School of Government.

Saving Arabic. (2014). Retrieved September 10, 2018, from http://www.lau.edu.lb/news-events/news/archive/saving_arabic

Saygin, A., Cicekli, I., & Akman, V. (2000). Turing Test: 50 Years Later. Minds Mach. 10, 4 (November 2000), pp. 463-518. DOI: https://doi.org/10.1023/A:1011288000451

Schubert, L. (2019). Computational Linguistics. The Stanford Encyclopedia of Philosophy. (Spring 2019 Edition), Edward N. Zalta (ed.), Retrieved February 21, 2019, from https://plato.stanford.edu/archives/spr2019/entries/computational-linguistics

Shoukry, A., & Rafea, A. (2012a). Sentence-Level Arabic Sentiment Analysis. 10.1109/CTS.2012.6261103.

Shoukry, A., & Rafea, A. (2012b). Preprocessing Egyptian Dialect Tweets for Sentiment Mining. In: 4th Workshop on Computational Approaches to Arabic Script-Based Languages, pp. 47–56.

Széll, M. (2011). Westernizing Arabic: Attempts to "simplify" the Arabic script. Tipográfiai diákkonferencia, Budapest, Hungary. Retrieved September 10, 2018, from http://emc.elte.hu/~hargitai/konferencia/szell.pdf

Tobaili, T. (2015). Sentiment analysis for Arabizi in social media (Master's thesis, Lebanese American University, 2015) (pp. 1-65). Beirut: LAU. DOI: https://doi.org/10.26756/th.2015.27

Turing, M. A. (1950). Computing Machinery and Intelligence. Mind, Volume LIX, Issue 236, 1 October 1950, pp. 433–460. DOI: https://doi.org/10.1093/mind/LIX.236.433

Turney, P. D., & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems (TOIS), 21(4), pp. 315–346.

Saving Arabic. (2014). LAU News. Retrieved from https://www.lau.edu.lb/news-events/news/archive/saving_arabic/

Valant, J. (2015). Online consumer reviews. European Parliamentary Research Service (EPRS), pp. 1-10. Retrieved April 30, 2019, from https://www.eesc.europa.eu/sites/default/files/resources/docs/online-consumer-reviews---the-case-of-misleading-or-fake-reviews.pdf

Versteegh, K. (1997). The Arabic Language. Edinburgh: Edinburgh University Press.

Wang, S., Chaovalitwongse, W., and Babuska, R. (2012). Machine Learning Algorithms in Bipedal Robot Control. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 42, no. 5, pp. 728-743. DOI: 10.1109/TSMCC.2012.2186565.

Waters, J. (2010). The Everything Guide to Social Media: All you need to know about participating in today's most popular online communities. Adams Media.

Wiebe, M., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2004). Learning subjective language. Computational Linguistics, 30: pp. 277–308.

Williams, C. (2014). Ukrainian teen created in lab passes Turing Test – famous nutty prof. Retrieved March 2, 2019, from https://www.theregister.co.uk/2014/06/09/software_passes_turing_test

Wright, A. (2009). Mining the Web for Feelings, Not Facts. Retrieved October 30, 2018, from https://www.nytimes.com/2009/08/24/technology/internet/24emotion.html

Yaghan, M. (2008). "Arabizi": A Contemporary Style of Arabic Slang. *Design Issues*, *24*(2), pp. 39-52. Retrieved September 24, 2018, from http://www.jstor.org/stable/25224166