People's Democratic Republic of Algeria

Ministry of Higher Education and Scientific Research

University of Laarbi Tbessi

Faculty of Exact Sciences and Natural and Life Sciences

Department of Mathematics and Computer Science

Final Study Dissertation

In the aim obtaining of MASTER Degree - Academic

Domain: Computer Science

Option: Information Systems

Presented By

Hamza GATTAL

Opinion mining in social networks: Algerian dialect as case study.

Date of presentation: 22-06-2021

In front of the jury

Ali Abdelatif BETOUIL	MCB	University of Laarbi Tbessi	President
Ala DJEDDAI	MCB	University of Laarbi Tbessi	Examiner
Hakim BENDJANNA	Pr	University of Laarbi Tbessi	Supervisor
Dalila BOURAS	Dr	University of Laarbi Tbessi	Co-Supervisor

Academic year: 2020/2021

Abstract

Sentiment analysis is a task of natural language processing which has recently attracted increasing attention. While significant work was directed toward the sentiment analysis of English text there is limited attention in literature toward the sentiment analytic of Arabic language and more especially Algerian dialect. This work focuses on the various supervised Sentiment Analysis methods available in the existing literature. Further, this work presents the comparative study of different Sentiment Analysis algorithms on the basis of accuracy.

Keywords: Opinion Mining, Sentiment Analysis, Arabic Language, Algerian Dialect.

Résumé

L’analyse des sentiments est une tâche de traitement du langage naturel qui a récemment attiré une attention croissante. Alors que d’importants travaux ont été dirigés vers l’analyse des sentiments du texte anglais, il y a une attention limitée dans la littérature vers le sentiment analytique de la langue arabe et plus particulièrement le dialecte algérien. Ce travail porte sur les diverses méthodes d’analyse des sentiments supervisées disponibles dans la littérature existante. En outre, ce travail présente l’étude comparative de différents algorithmes d’analyse de sentiment sur la base de la précision.

Mot cles: Fouille des Opinions, Analyse des Sentiments , Langue Arab, Dialecte algérien.

ملخص

اجتذب مؤخرا تحليل المشاعر اهتماما متزايدا. حيث أنه تم توجيه عمل كبير نحو تحليل المشاعر للنص باللغة الإنجليزية، هناك اهتمام محدود نحو تحليل المشاعر في اللغة العربية وبالأخص اللهجة الجزائرية. يركز هذا العمل على أساليب تحليل المشاعر في المؤلفات الموجودة في عمل سابق. وعلاوة على ذلك ، يعرض هذا العمل الدراسة المقارنة لخوارزميات تحليل المشاعر المختلفة على أساس الدقة.

الكلمات المفتاحية : تحليل المشاعر ، اللغة العربية ، اللهجة الجزائرية.

Thanks

First, we would like to express our gratitude to Allah for giving us with the determination, will, and energy to finish this modest work.

We thank our supervisor and co-supervisor Pr. Hakim BENDJANNA and Dr. Dalila BOURAS, for their guidance, patience, availability, and support throughout this period.

We would like to express our heartfelt gratitude and sincere appreciation to:

Jury members for agreeing to judge our work and to enrich it. All the people who helped us, from near or far.

We also extend our thanks to all our teachers who have watched over our training.

List of Figures

List of Tables

Contents

General Introduction

General Introduction

Currently, our lives are based on information and its analysis. This information is more available today and more precisely in digital form, with the development of Web 2.0. More and more, people communicate, share content and express their opinions on the Internet about a wide range of topics, in newsgroups, blogs, forums and other sites regarding product reviews.

The social networks platforms have a large number of users, and that they share their views everyday on products, ideas, services, etc. Despite the massive amount of comments and reviews, the opinion mining (OM) or sentiment analysis (SA) made it easier to access useful information.

Day after day, the importance of opinion mining is getting bigger, especially in businesses and marketing, It can give a thought of what people like and do not like.

The opinion mining field may be a classification of a choice of opinion as positive, negative or neutral. Its main purpose is to extract users opinions from social networks, for instance , using automated techniques to define their positions on an issue , which is usually expressed in text form.

The English language is receiving considerable interest from researchers within the field of opinion mining or sentiment analysis [Pang and Lee, 2008].

In contrast to the Arabic language, where there are a limited number of researches. Most of them focus on Modern Standard Arabic (MSA) [Duwairi et al., 2014], among which few number focus on Arabic dialects like Tunisian Dialect [Medhaffar et al., 2017], Saudi Dialect [Al-Twairesh et al., 2018], Jordanian Dialect [Atoum and Nouman, 2019], Algerian Dialect [Mataoui et al., 2016] and more.

The purpose of this work is to study the general public opinions which concern essentially the sentiments found in the Facebook comments written in Arabic and more specifically in Algerian dialect.

Our work is split into two main chapters. In the first chapter, we focus on the state of the art, we describe the opinion mining and present the Arabic language and the Algerian dialect. The second chapter, we present intimately the modelling of our system and discuss about the results obtained.

Chapter 1

State of the Art

Chapter 1

State of the Art

1.1 Introduction

Social media may be a tool that's becoming quite popular lately due to its user-friendly features. Its platforms like Facebook, Instagram, Twitter and others are giving people an opportunity to communicate to one another. Subsequently, different people have different opinions on this debatable topic.

The newer research on language analysis in social media has been increasingly that specialize in the latter's impact on our daily lives. Opinion mining is one among the foremost promising avenues for data processing. It's a scientific challenge to develop methods and algorithms which extract information from data coming from multiple sources and languages in various formats.

1.2 Opinion Mining

1.2.1 Definition

According to [Liu, 2012], sentiment analysis also called opinion mining, is that the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities like products, services, organizations, individuals, issues, events, topics, and their attributes. It represents an outsized problem space.

Opinion Mining is taking opinions from different sources like online shopping sites, review sites, blogs, social networking sites like Facebook and Twitter, news portals, etc., on a particular area or problem[Seerat and Azam, 2012].

Opinion Mining might be a way of assembling reviews/feedback's which are shared on-line on various social media websites, blogs and e-commerce internet sites followed by the extracting information within the interior data.

1.2.2 Opinion Mining Levels

In general opinion mining or sentiment analysis has been investigated mainly at three levels, it can be done word level, sentence level and document level.

Document Level

[Pang et al., 2002] In the document level all opinions contained in all documents Classified on the basis of categories (as positive or negative). For example the document level task determine whether the review given to a product is positive or negative.

Sentence Level

At the sentence level, digging deeper into the documents from previous level, each sentence has been categorized into a category (positive, negative, neutral).

Aspect Level

The previous levels consists of a sentiment such as positive or negative. The aspect level (or feature level) looks directly at the opinion itself, not only at the sentiment such as positive or negative, it also look at its target. opinion targets are explained with their different aspects.

1.2.3 Application of Opinion Mining

Opinion mining have been considered by many companies as a part of their mission, it have a important role in imparting sub-component technology for other systems. Applying sentiment analysis on the feedback's of individuals provides valuable information for further analysis on market reports. The main application of opinion mining and sentiment analysis are

- Policy Making and Decision Making
- Buying Products or Services
- Recommendation Systems
- Quality Improvement
- Marketing research

1.2.4 Sentiment Analysis Approach's

Sentiment analysis has been practiced on a variety of topics. For instance, sentiment analysis studies for movie and product reviews, and news and blogs [Thakkar and Patel, 2015].

The sentiment analysis approach's divided to three techniques, which are Machine Learning based approach, Lexicon based approach and Hybrid/Combined approach [Alhojely, 2016].

Machine Learning Based Approach

Machine learning is one among the foremost prominent techniques gaining interest of researchers. in sentiment analysis, the classification methods that use Machine Learning based approach divided into supervised and unsupervised learning method [Alhojely, 2016].

Machine Learning based approach comprises of three stages:

- Data collection
- Data prepossessing
- Data training

Lexicon Based Approach

Lexicon based approach is an approach that uses a dictionary and contains polarity of the word in it. If a word appears in a text, it will be compared with a word in the dictionary, and the sentiment score are going to be added. The determination of the sentiment is using the lexicon-based approach, then it is calculated by the total of the polarity contained in a text [Isabelle et al., 2019].

Hybrid/Combined approach

Hybrid/Combined approach is a combination of previous approach's Machine Learning based approach and Lexicon based approach.

1.3 The Arabic Language and Algerian Dialect

Millions of people use social media platforms like Facebook, Twitter , Instagram and more. Everyone uses one or many languages to write their topics. In the Arab world, most users of social media use the Arabic language to disseminate and comment on topics, These users rely on the modern standard Arabic or their local dialects.

1.3.1 The Arabic Language

The Arabic language is a Semitic language spoken in in a large area including North Africa, most of the Arabian Peninsula (Jazīrat Al-'Arab), and other parts of the Middle East As shown in figure 1.1. Arabic is the language of the Qur'ān.

Modern Standard Arabic is the official language of the Arab World. It is the primary language of the media and education.

The Arabic dialects are the true native language forms. They are generally restricted in use for informal daily communication.

1.3.2 Modern Standard Arabic

The Modern Standard Arabic (MSA), or Modern Written Arabic, is the written form of the Arabic language, differs in a non-trivial fashion from the various spoken varieties of Arabic [Zaidan and Callison-Burch, 2014]. MSA is a term used mostly by Western linguists to refer to the variety of standardized, literary Arabic that developed in the Arab world in the late 19th and early 20th centuries. It is the official language used in academia, print and mass media, law and legislation, though it is generally not spoken as a mother tongue [Habash, 2010]. MSA is much more modern. MSA is primarily written not spoken.

Figure 1.1: Map of the Arab world

1.3.3 The Arabic dialects

The Arabic dialects also called colloquial Arabic, in contrast, are the true native language forms. They are generally restricted in use for informal daily communication.They are not taught in schools or even standardized although there is a rich popular dialect culture of folk-tales, songs, movies, and TV shows. Dialects are primarily spoken not written [Habash, 2010].

[Habash, 2010] and [Versteegh, 2014] give a breakdown for spoken Arabic dialects into groups:

- Egyptian Arabic: covers the dialects of Egypt and Sudan.
- Levantine Arabic: covers the dialects of Lebanon, Syria, Jordan and Palestine.
- Gulf Arabic: covers the dialects of Kuwait, United Arab Emirates,Saudi Arabia, Oman, Bahrain, and Qatar.
- North African (Maghrebi) covers the dialects of Morocco, Algeria, Libyan, Tunisia and Mauritania.
- Iraqi Arabic.
- Yemenite Arabic.

1.3.4 The Arabic Script

The Arabic script is used to write Arabic language, there are also many languages around the world like Persian, Kurdish are written using Arabic script. Arabic dialects are by default written in Arabic script although there are no standard dialectal spelling systems [Habash, 2010]. In the Arabic script two types of symboles for writting are: letters (Figure 1.2) and diacritics (Figure 1.3). writtes from right to left.

[Habash et al., 2012] define give a brief definition of letters and diacritics as follows: Arabic letters are written in cursive style in both print and script (handwriting). Diacritics are additional zero-width symbols that appear above or below the letters.

The basic 28 letters	
ا ب ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك ل م ن ه و ي	
The Hamza letters	
أ إ آ ؤ ئ ء	
The Ta-Marbuta	**The Alif-Maqsura**
ة	ى

Figure 1.2: The Arabic script letters

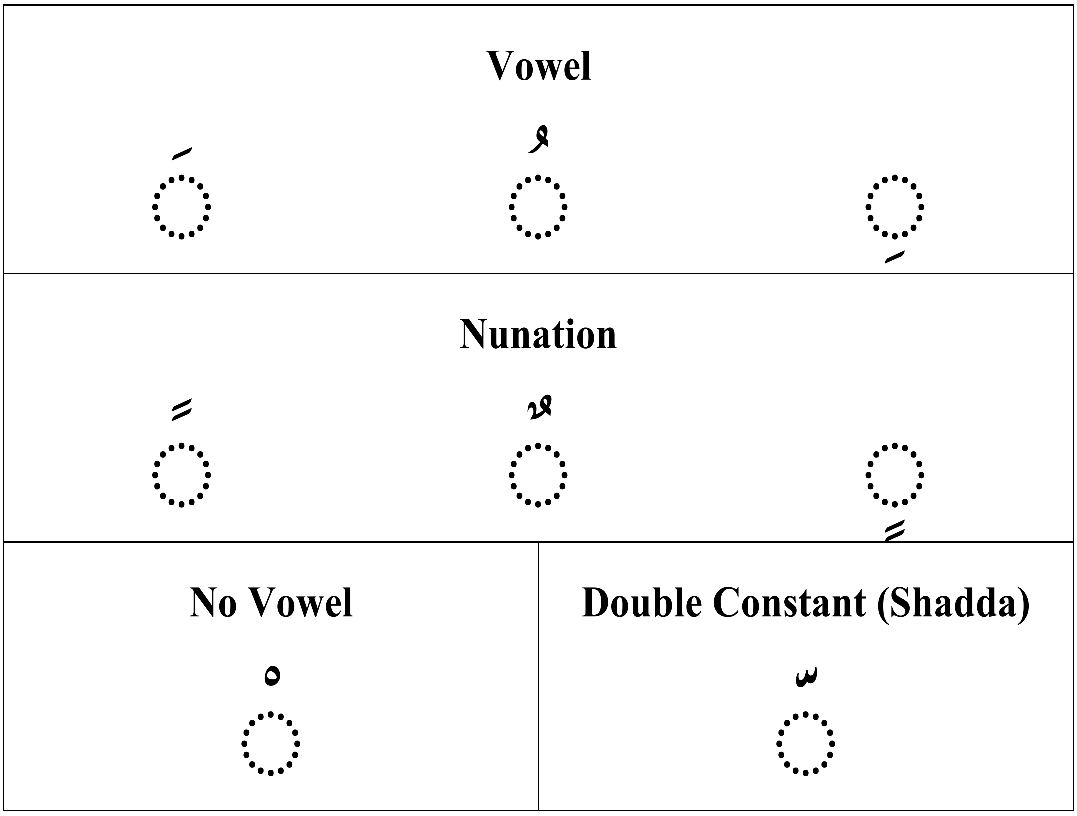

Figure 1.3: The Arabic script diacritics

1.3.5 Algerian dialect

The Algerian dialect or Algerian Arabic is a group of North African Arabic dialects (Maghrebi) mixed with different languages spoken in Algeria [Adouane and Dobnik, 2017]. Algerian Arabic has a lot of influences from Berber as well as French, Classical Arabic, Modern Standard Arabic, and English [Habash, 2010].

1.4 The Arabic and Opinion Mining Challenges

1.4.1 Arabic and Dialects Challenges

The Arabic as Modern Standard Arabic or Arabic dialects poses many challenges

- A single dialect (for example the Algerian dialect) may contain several sub-dialects.
- There's a big difference between Modern Standard Arabic and Arabic dialects.

- A root of a word can take many forms depending on the context.
- The repetition of a letter several times to intensify the sense or feeling.
- The presence or absence of diacritics, can completely change the meaning of words.
- Words of negation used to deny verbs in the past or in the present, which change the meaning exactly to the contrary.

1.4.2 Challenges in Opinion Mining

The extraction of sentiment consists in determining the polarity of an opinion, it is may be positive, negative or neutral. For example good, fine, amazing, wonderful are positive sentiments, and bad, ugly, poor, terrible are negative sentiments. Although sentiment words and phrases are important for sentiment analysis there are several issues related to usage of these opinion words.

An opinion word may have different orientations in different applications. For example “This movie has an unpredictable plot” is a positive phrase, while “Your vehicle has an unpredictable steering wheel” is a negative one. The opinion word unpredictable is used in different ways.

Sarcastic statements are much harder to deal. for example “What a wonderful camera! It stopped working in two days” Here the word wonderful means a positive opinion but actually the statement made is negative.

Opinion spamming problem in sentiment analysis. People post fake opinions to promote or to discredit products, services, organizations etc.

1.5 Related work

There are several Research related to Arabic sentiment analysis field with focus on dialectical Arabic study cases.

Arabic language is characterized by a wide number of dialects varieties. Besides Modern Standard Arabic used as a formal language, different Arabic dialects are used for nearly all everyday speaking situations. By the emergence of social media and the various electronic networks, enabling Arab users to express their opinions using different Arabic dialects, researchers have raised the need to consider this amount of generated content especially by the study of the peculiarities related to written forms of these different dialects [Mataoui et al., 2016].

1.5.1 Case of Modern Standard Arabic

The work of [Ibrahim et al., 2015] presents a feature-based sentence level approach for Arabic sentiment analysis. This approach is a semi-supervised approach for sentiment analysis used Arabic idioms/saying phrases lexicon to detect the sentiment polarity in Arabic sentences.

They built their own corpus contains 2000 Arabic sentiment statements includes 1000 MSA tweets, Arabic dialect tweets and 1000 microblogs. The corpus includes data in both Modern Standard Arabic and Egyptian dialectal Arabic. They collect about 10 thousand Arabic tweets and 10 thousand Arabic comments and reviews. The selected data is performed according to specific conditions. The data is selected and annotated manually as; positive and negative sentiment. They built two lexicons; Arabic sentiment words lexicon and Arabic sentiment idioms/saying phrase lexicon. Most of the work use adjectives only for sentiment analysis, and some of them use nouns, verbs, adverbs or a combination of them. They introduce a (5244) sentiment adjectives lexicon “ArSeLEX“ which is manually created and automatically expandable. Starting with the gold-standard 400 adjectives collected manually as a seed from different websites specialist in Arabic language and grammar. The lexicon is expanded manually for the first time by collecting synonyms and antonyms of each word using different Arabic dictionaries and label each word with one of the following tags [negative (NG), positive (PO), neutral (NU)].

The system consists of two parts; part1 includes the pre-processing and lexicon expansion, part2 includes the features extraction and classification. The pre-processing includes data cleaning, remove stop words and normalization. The features extraction and classification includes applying standard features, sentence level features, linguistic features and Syntactic features for conflicting phrases.

They use the Support Vector Machines (SVM) for classification. and run two sets of experiments. In the first set they divide the data into 80% for training and 10% for developing and 10% for testing. The classifier and features are optimized during the developing set and all the results that we report are in the test set, and report accuracy, precision, recall and F-measure.

Their corpus consists of 2000 topics that include tweets, hotel reservation comments, product reviews and TV-Programs comments. Each topic labeled as negative or positive.

As shown in the tables 1.1,1.2,1.3 and 1.4 the results obtained show that their sentiment analysis system is very promising, it show that their sentiment analysis system yields high performance and efficiency in sentiment classification of the types of data that applied to the system.

Table 1.1: Baseline results for Arabic sentiment classification before lexicon expansion

Data	Accuracy	Precision	Recall	F-Measure
Tweets	80.95238%	84.94624%	84.94624%	84.9462%
Hotel reservation comments	94.63087%	98.26087%	94.95798%	96.5812%
Product reviews	93.84615%	96.22642%	96.22642%	96.2264%
TV-Programs comments	88.40580%	92.10526%	87.50000%	89.7436%
Total data	89.06977%	93.31104%	91.17647%	92.2314%

Table 1.2: Baseline results for Arabic sentiment classification after lexicon expansion

Data	Accuracy	Precision	Recall	F-Measure
Tweets	83.67347%	87.09677%	87.09677%	87.0968
Hotel reservation comments	95.30201%	98.27586%	95.79832%	97.0213
Product reviews	95.38462%	98.11321%	96.29630%	97.1963
TV-Programs comments	94.20290%	97.43590%	92.68293%	95.0000
Total data	90.46512%	94.31438%	92.15686%	93.2231

Table 1.3: Results for Arabic sentiment classification before lexicon expansion

Data	Accuracy	Precision	Recall	F-Measure
Tweets	87.50000%	84.1804%	100%	91.41085
Hotel reservation comments	93.54839%	95.83333%	95.83333%	95.83333
Product reviews	96.66667%	95.65217%	100%	97.77777
TV-Programs comments	93.33333%	93.33333%	93.33333%	93.33333
Total data	94.30894%	91.78082%	98.52941%	95.03545

Table 1.4: Results for Arabic sentiment classification after lexicon expansion

Data	Accuracy	Precision	Recall	F-Measure
Tweets	90.62500%	88.00000%	100%	93.61702
Hotel reservation comments	96.77419%	96.00000%	100%	97.95918
Product reviews	100%	100%	100%	100
TV-Programs comments	96.66667%	100%	93.33333%	96.55172
Total data	95.12195%	93.15068%	98.55072%	95.77464

Results in tables 1.2 and 1.4 shows the effect of lexicon expansion on sentiment classification. Also, Tables 1.1 and 1.3 reported the gradual increases in the polarity lexicon coverage.

1.5.2 Case of Tunisian Dialect

The work of [Medhaffar et al., 2017] focus on SA of the Tunisian dialect. They use Machine Learning techniques to determine the polarity of comments written in Tunisian dialect.

They start by using and evaluating the performance using available resources from Modern Standard Arabic and dialects, then create and annotate their own data set.

The corpus called TSAC (Tunisian Sentiment Analysis Corpus), is collected from Facebook users comments in popular pages in Tunisia. The collected corpus contains 17k user comments manually annotated to positive and negative polarities.

Table 1.5: Statistics of the TSAC corpus

	Positive	Negative	Total
Comments	8215	8845	17060

In data training, they used three 3 different training corpus, [Bayoudhi et al., 2015] OCA (Opinion Corpus for Arabic), [Aly and Atiya, 2013] LABR (Large-scale Arabic Book Review) and TSAC. For classification they used machine learning methods are Support Vector Machines (SVM), Naive Bayes and Multi-Layer Perceptron (MLP).

The results of the different classifiers are presented: the error rate of 0.23 with SVM, 0.22 with MLP and 0.42 with NB. SVM and MLP obtain similar results. However, lower results are obtained with NB classifier.

1.5.3 Case of Jordanian Dialect

The work of [Atoum and Nouman, 2019] proposes a sentiment analysis model of Arabic Jordanian dialect tweets.

They proposed a model analyzes, mines, and classifies Arabic Jordanian dialect tweets. The model consists 4 phrases are: Collecting Tweets, Tweets Extraction, Cleaning and Tweets Annotations and Tweets Preprocessing.

The collecting tweets phrase is to collect a corpus of Jordanian dialect tweets. The tweets extraction phrase is to extract the important content from the tweet. The cleaning and tweets annotations phrase is to remove links and some special symbols from the collected tweets. The tweets preprocessing phrase consists 6 stages are:

- Cleaning Stage: each tweet contains a special symbols and various characters such as emoticons take a new classification.
- Normalisation: remove all extra spaces and replace any un-normalized letter by its normalized form.

- Tokenization
- Named Entity Recognition
- Removing stop words
- Stemming: removing any attached suffixes, prefixes, and/or infixes from words in tweets.

For classification, They used the Support Vector Machines and Naïve Bayes machine learning algorithms. To compare the performance of two classifiers to compare the performance of Naïve Bayes and Support Vector Machines classifiers, They conducted several experiments. The results obtained from conducting these experiments are shown the Support Vector Machines classifier performs better than the Naïve Bayes.

Table 1.6: SVM Experimental Results

Measure	Precision	Recall	F-measure	Roc-Area	Accuracy
1st Experiment	0.85	0.84	0.84	0.87	0.84
2nd Experiment	0.83	0.83	0.83	0.82	0.82
3rd Experiment	0.75	0.75	0.75	0.80	0.75
4th Experiment	0.77	0.74	0.74	0.78	0.74
5th Experiment	0.82	0.76	0.75	0.79	0.76
6th Experiment	0.73	0.73	0.72	0.78	0.73

1.5.4 Case of Algerian Dialect

The work of [Abdelli et al., 2019] presents supervised method for sentiment analysis of Arabic Algerian dialect. They apply two supervised methods on huge annotated data set.

They approach consists of 3 phases:

- Data collection: collecting a large data set from different Arabic Algerian sources, and annotate them.

Table 1.7: NB Experimental Results

Measure	Precision	Recall	F-measure	Roc-Area	Accuracy
1st Experiment	0.53	0.53	0.53	0.70	0.53
2nd Experiment	0.55	0.54	0.54	0.70	0.54
3rd Experiment	0.49	0.47	0.47	0.64	0.47
4th Experiment	0.64	0.55	0.50	0.74	0.55
5th Experiment	0.68	0.55	0.50	0.86	0.55
6th Experiment	0.51	0.50	0.50	0.67	0.50

- Data preprocessing: preprocess the collected data.
- Data training: train and test the two models.

From popular Algerian Facebook pages, they collected more than 100K comments. and labeled more than 10K comments into positive and negative comments. They also used other data sets, then combined then in one huge data set. For the Word2Vec, they collected a big text corpus of 1.4 gigabytes.

Table 1.8 show the results of the Support Vector Machines model with Tf-Idf techniques and The Long Short-Term Memory model with Word2Vec(CBOW).

Table 1.8: SVM & LSTM results

Classifier	Accuracy	Precision	Recall	F1-Score
SVM	0.86	0.89	0.82	0.85
LSTM	0.81	0.79	0.75	0.77

After trained the two models. The Support Vector Machines model outperform The Long Short-Term Memory model.

1.5.5 Case of Saudi Dialect

The work of [Al-Twairesh et al., 2018] develop a hybrid method for sentiment analysis for Arabic tweets for Saudi Dialect.

They engineered and evaluated several features using the feature backward selection method. Then developed a hybrid method for classification models. The proposed method uses a set of dialect independent features and a large corpus of Saudi tweets. with a lexicon-based method. As shown in the table 1.9 the feature set contains a 19 feature divided into a subset (Semantic, Stylistic and Tweet Specific).

Table 1.9: Features used in classification model

Subset	Features	
Semantic	hasPositiveWordAraSenTi	hasNegativeWordAraSenTi
	hasPositiveWordMPQA	hasNegaitveWordMPQA
	hasPositiveWordLiu	hasNegaitveWordLiu
	hasNegation	hasIntensifier
	hasDiminisher	hasModalWord
	hasContrastWord	PositiveWordCount
	NegativeWordCount	TweetScore
Stylistic	hasQuestionMark	hasExclamationMark
	hasPositiveEmoticon	hasNegativeEmoticon
Tweet Specific	tweetLength	

They developed and compared three models for classification (Two-way, Three-way and Four-way). The classification is performed using a 15 label.

The corpus consists of tweets written in Modern Standard Arabic and the Saudi Dialect, it is the largest corpus of Saudi tweets [Al-Twairesh et al., 2017]. It contains more than 17K tweets labeled by four labels (positive, negative, neutral, and mixed).

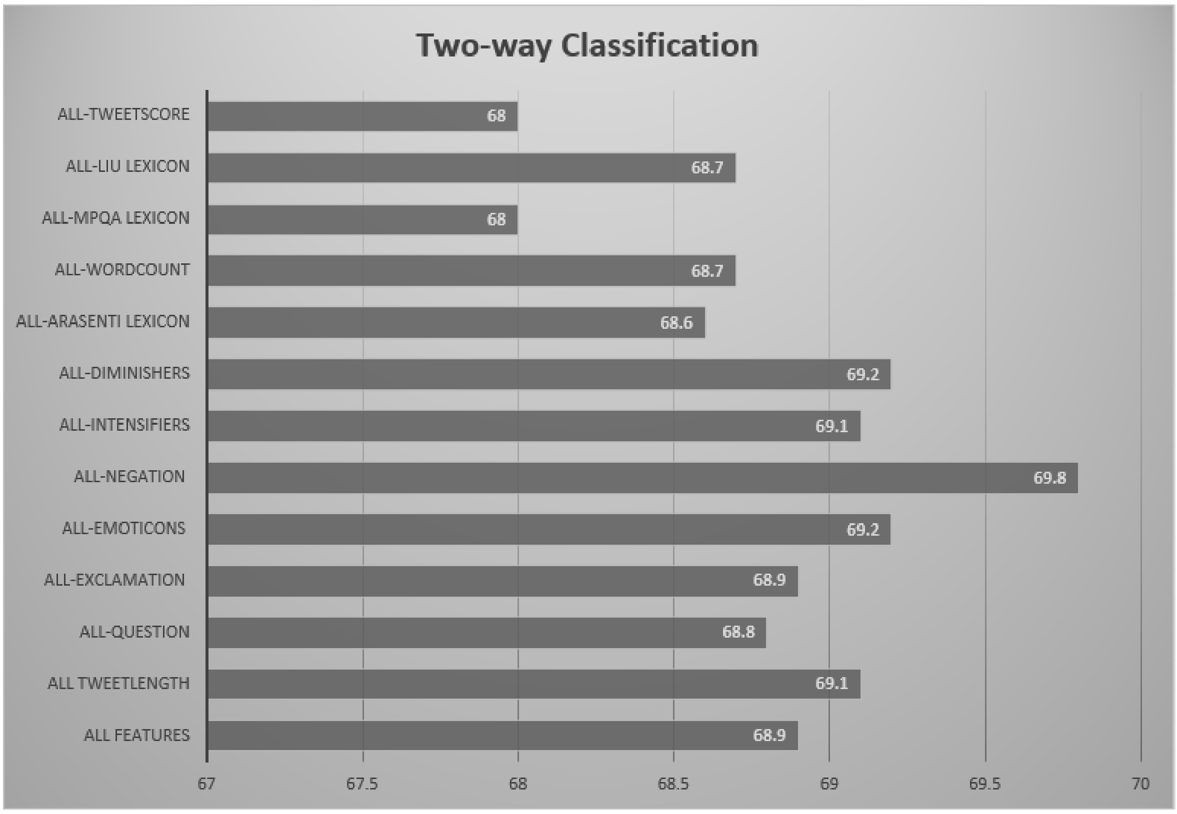

Figure 1.4: F1-Score for Two-way classification model [Al-Twairesh et al., 2018]

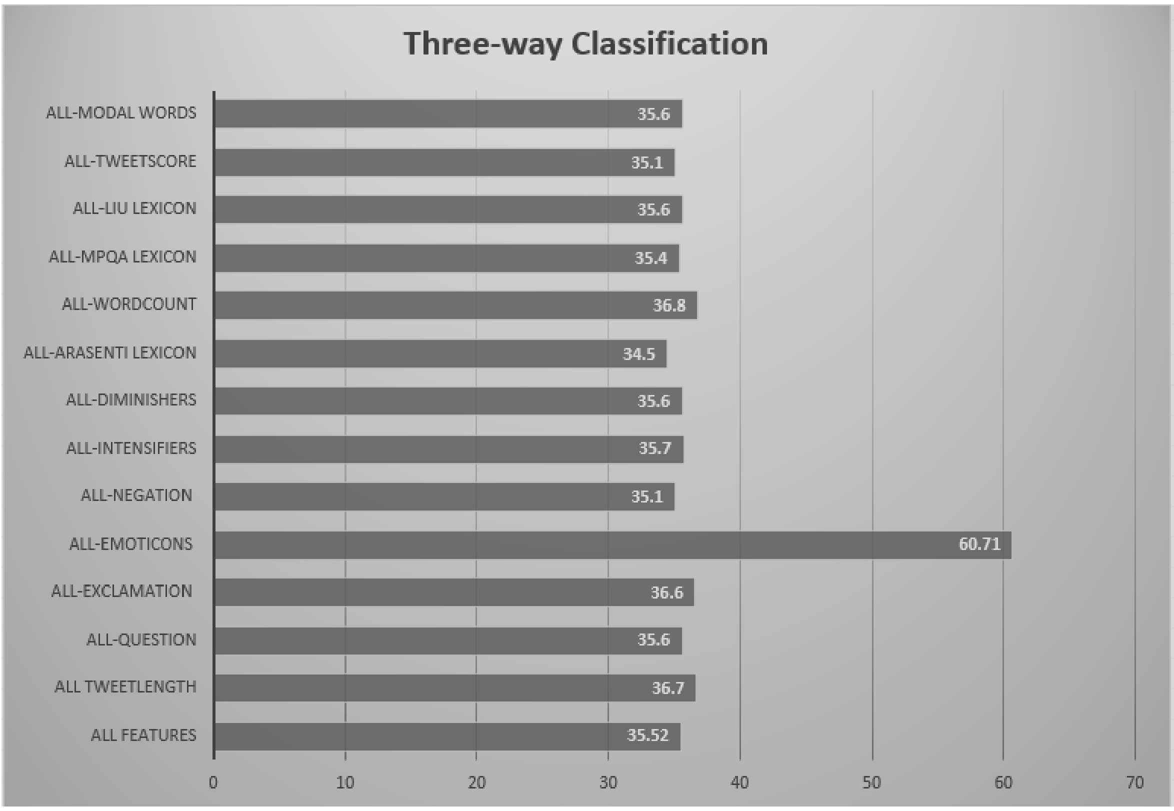

Figure 1.5: F1-Score for Three-way classification model [Al-Twairesh et al., 2018]

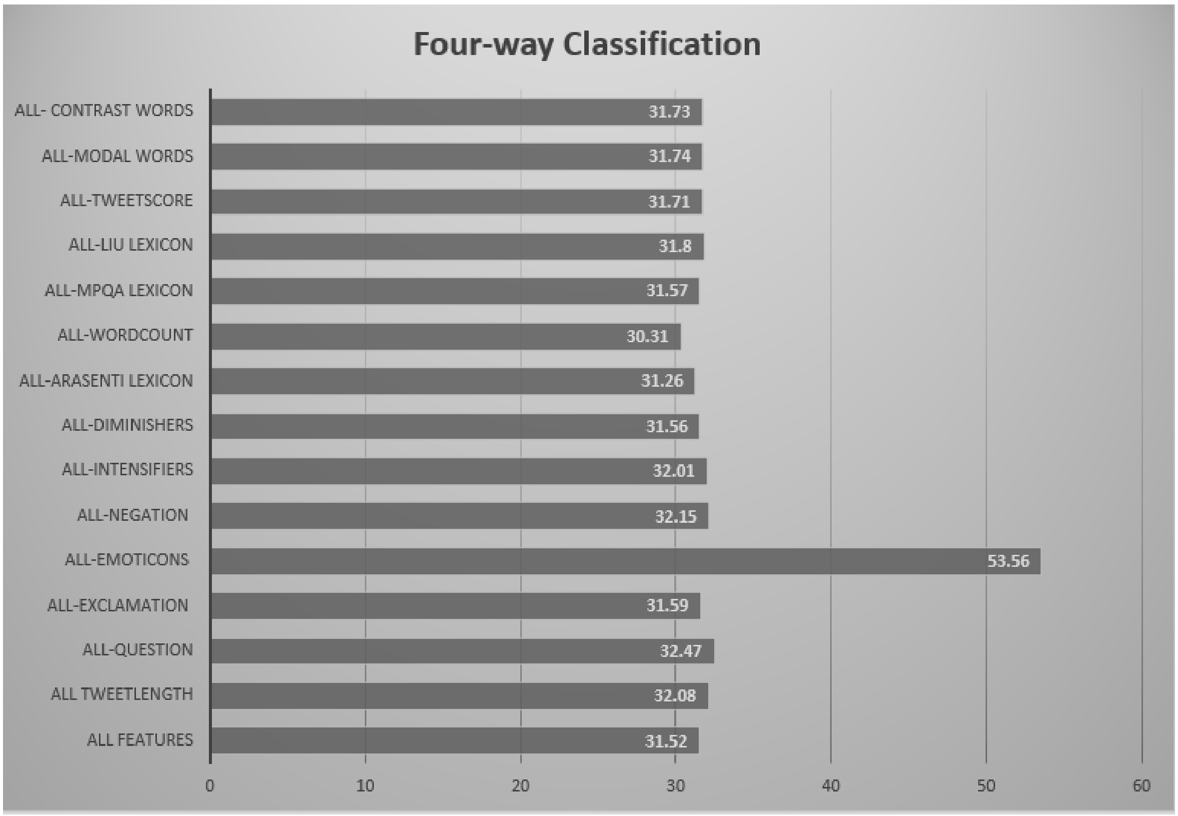

Figure 1.6: F1-Score for Four-way classification model [Al-Twairesh et al., 2018]

In Two-way classification model all features included except two features (asModalWord and hasContrastWord) used to identify neutral and mixed classes. In Three-way classification model all features included except a feature (hasContrastWord) used to identify the mixed class. In Three-way classification model all features included. The Figures 1.4, 1.5 and 1.6 shows the F1-Scores after removing each feature.

In the Two-way classification model 6 features increase the F1-Score. after removing features (remove one feature and recalculate the F1-Score) the F1-Score increased. In the Three-way classification model 12 features increase the F1-Score. after removing features the F1-Score reached the 61.5%. In the Three-way classification model 15 features increase the F1-Score. After reduced feature set and the F1-Score was calculated for the classifier and it was 55.07%.

1.6 Conclusion

In this chapter, we have studied the opinion mining and the Arabic language. First, we have detailed the opinion mining, its levels, its approach's and presented the application domains. Next, we have presented the Arabic language, Modern Standard Arabic, Arabic script and Algerian dialect. Then, we have defined the challenges of opinion mining and Arabic language and its dialects. Finally, we have presented some related works in Arabic language and its dialects.

In next chapter, six classifiers would be used to analyze sentiment.

Chapter 2

Proposed Approach

Chapter 2

Proposed Approach

2.1 Introduction

In this chapter, we begin to create our opinion analysis model. In starting with our contribution. First, we describe the data source and the data preprocessing steps. Next, we will mainly implement four algorithms. These algorithms are: Support Vector Machines (SVM), Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB) and Logistic Regression (LR). Finally, we will discuss the results of the classification.

2.2 Contribution

Our main contributions is a work on the Algerian dialect, with the operation of six classifiers: Support Vector Machines (SVM), Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), Logistic Regression (LR) and Long Short-Term Memory (LSTM), which is considered to be our work that tests the use of the six classifiers.

2.3 Data source

To save time, we exploited the dataset used in the work of [Abdelli et al., 2019]. They built their own dataset in Algerian dialect.

The dataset contains:

- LABR dataset [Aly and Atiya, 2013].

- Multi domaine dataset of [ElSahar and El-Beltagy, 2015].
- Dataset of Algerian Arabic comments of [Mataoui et al., 2016].
- 10k comments labeled into positive and negative, collected from popular Algerian Facebook pages.

The LABR is a sentiment analysis dataset for the Arabic language. It consists of over 63,000 book reviews, each rated on a scale of 1 to 5 stars. It can be used for two tasks: sentiment polarity classification and rating classification.

The Multi domaine dataset consist of a total of 33K annotated reviews for movies, hotels, restaurants and products.

After deleting the duplicate items, and after balancing the all the dataset as shown in table 2.1 contains 49,864 comments that are evenly divided between positive and negative comments.

Table 2.1: Dataset statistics

	Positive	Negative
Total comments	24932	24932
Total words	1180663	1345029
Average words in each comment	47.36	53.95
Average characters in each comment	253.15	294.47

From the same work of [Abdelli et al., 2019], we used the text corpus for word embedding. this corpus consists of five corpora:

- The dataset.
- The Arabic Wikipedia corpus.
- The Open Source Arabic Corpora (OSAC) [Saad and Ashour, 2010].
- The Tashkeela corpus [Zerrouki and Balla, 2017].

- Corpus collected from Algerian news websites

All items of the dataset have been concatenated to form the first corpus.

2.4 Data preprocessing

The first phase in the preprocessing process was to remove non-Arabic content, followed by the normalization of Arabic characters.

According to [Abdelli et al., 2019], they deleted numbers and non-Arabic letters before normalizing some letters, as illustrated in Figure 2.1.

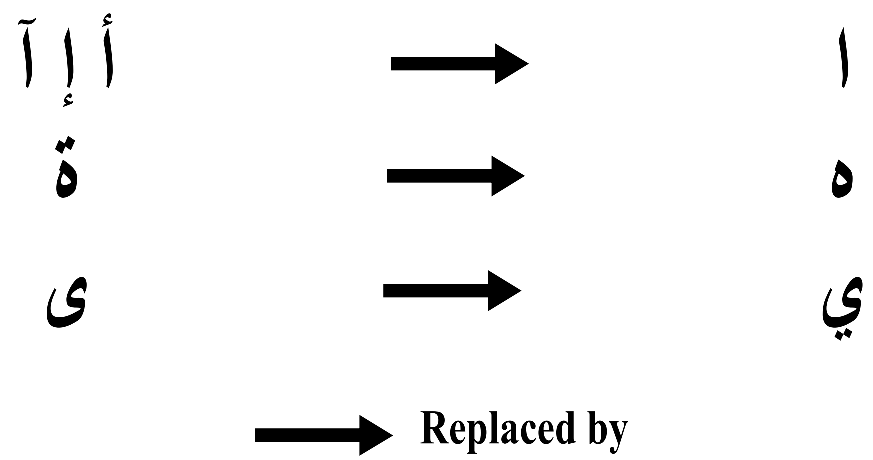

Figure 2.1: Normalized letters

We've listed some examples, along with their polarities, in Table 2.2.

2.5 Models training

We separated the dataset into a training set and a test set to train six models, allocating 80% of the dataset to training and 20% to testing. This study was conducted out using Python 3.7 on an HP notebook with an Intel(R) Core i3-6006U processor and 4 GB of RAM.

We used the scikit-learn framework to train SVM, RF, DT, NB, and LR models (sklearn). We employed the TF-IDF approaches with the scikit-learn function TfidfVectorizer for word

Table 2.2: Example of some comments

Comment	Polarity
اكب سعدك ممصتك اوه يا كذاب يا سفيه ملا كوارث والله	Negative
راك كبير علي هاذ التفاهات ا حفيظ	Negative
مكانش الزهر يروح يرقي	Negative
هشام الخلصي يا معذبهم والله معلم نقطه وارجع لسطر	Positive
يا حفيظ تطولت هاغيبه ربي يحفظك	Positive
صوت رائع و ماتيار برميار اروع	Positive

embeddings.

Scikit-learn is a Python module that integrates a wide range of cutting-edge machine learning methods for supervised and unsupervised medium-scale issues. [Pedregosa et al., 2011].

We utilized TensorFlow for LSTM training. We used the Gensim library's Word2Vec approach for word embeddings.

TensorFlow is a machine learning library that is free and open-source. It can be used for a variety of applications, but it focuses on deep neural network training and inference.

Gensim is an open-source library that uses modern statistical machine learning to perform unsupervised topic modeling and natural language processing.

We'll define each model, as well as TF-IDF and Word2Vec, briefly below:

2.5.1 Word Embeddings

Word embeddings, also known as word representations or word encoding, is a technique for mapping words to real-number vectors. Many approaches exist for word embeddings, such as Word2vec and Term Frequency-Inverse Document Frequency (TF-IDF).

The term frequency (TF) and inverse document frequency (IDF) are calculated in the TF-IDF (IDF). There are various ways to compute TF and IDF, but the equations below are the most commonly used:

$$tf(t,d) = \sum_{t' \in d} f_{t',d} \tag{2.1}$$

$$idf(t,D) = logN/n_t \tag{2.2}$$

t' is the term t occurrence in document d, n_t is the number of documents contains t and N is the number of documents.

The Word2vec was created by [Mikolov et al., 2013], and it is a method that represents each different word with a specific set of integers known as a vector. It has two models:

- The Continuous Bag-of-Words Model (CBOW).
- The Continuous Skip-gram Model.

2.5.2 Support Vector Machines

Support Vector Machines (also known as support-vector networks) are supervised machine learning models that employ classification methods to solve problems involving two groups. [Cortes and Vapnik, 1995] developed it at AT&T Bell Laboratories. Support Vector Machines are a type of machine learning algorithm that can be used for pattern identification and regression. It are based on statistical learning theory.

Mathematically, if the training data is linearly separable, then a pair (w, b) exists, such as $w^T x_i + b \geq 1$ for all $x_i \in P$ and $w^T x_i + b \leq -1$ for all $x_i \in P$, with the decision rule $f_{w,b}(n) = sgn(w^T x_i + b)$ for all (x i in P), where w is termed as weighted vector and b as the bias. It is simple to show that when two classes can be separated linearly, the best separating hyperplane may be obtained by reducing the separating hyperplane's squared norm [Malik and Mumtaz, 2019]. A convex quadratic programming (QP) problem can be used to solve the challenge of minimization: Minimize $\Phi(w) = \frac{1}{2}\|w\|^2$ subject to $y(wx + b) \geq 1$, $i = 1, 2, ..., l$ [Malik and Mumtaz, 2019].

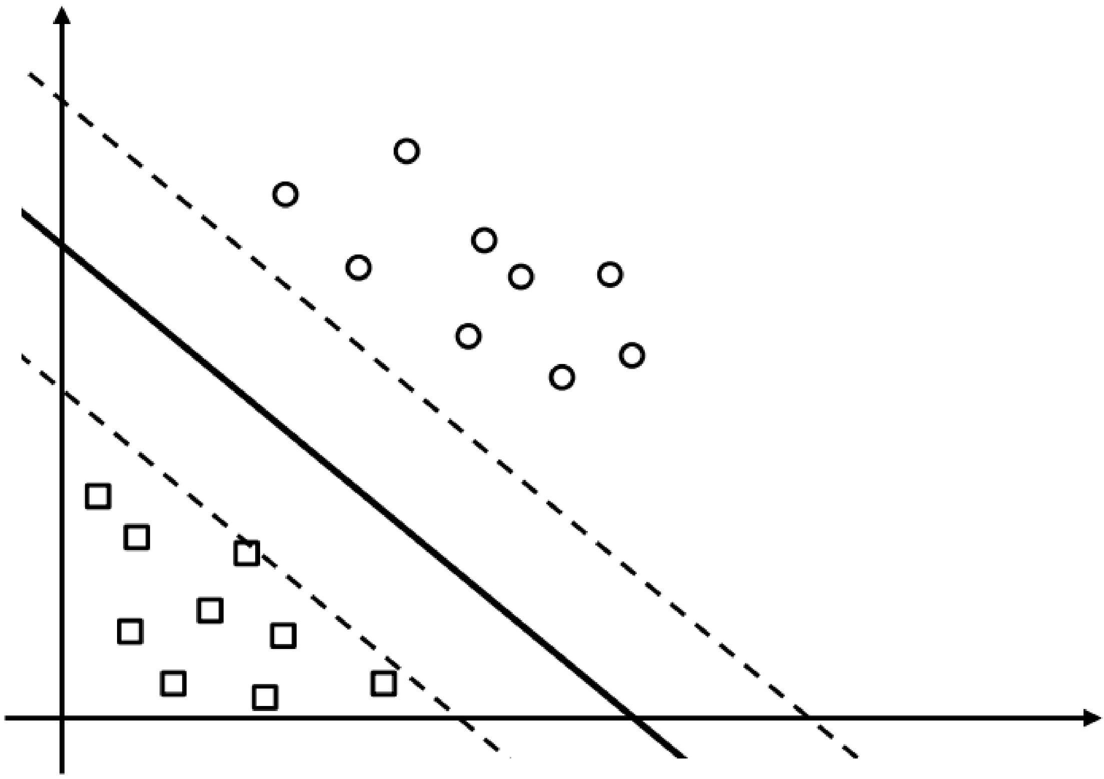

Figure 2.2: SVM trained with samples from two classes

2.5.3 Naive Bayes

The Naïve Bayes represents a supervised learning method as well as a statistical method for classification.

Naïve Bayes algorithm is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. Naïve Bayes provides a mechanism for using the information in sample data to estimate the posterior probability $P(y \mid x)$ of each class y, given an object x. Naïve Bayes is based on:

$$P(y \mid x) = P(y)P(x \mid y)/P(x) \tag{2.3}$$

$$P(x \mid y) = \prod_{i=1}^{n} P(x_i \mid y) \tag{2.4}$$

$$P(x) = \prod_{i=1}^{k} P(c_i)P(x \mid c_i) \tag{2.5}$$

2.5.4 Decision Trees

For classification and regression, decision trees are a non-parametric supervised learning method. The goal is to learn simple decision rules from data attributes to develop a model that predicts the value of a target variable. A tree is an approximation to a piecewise constant.

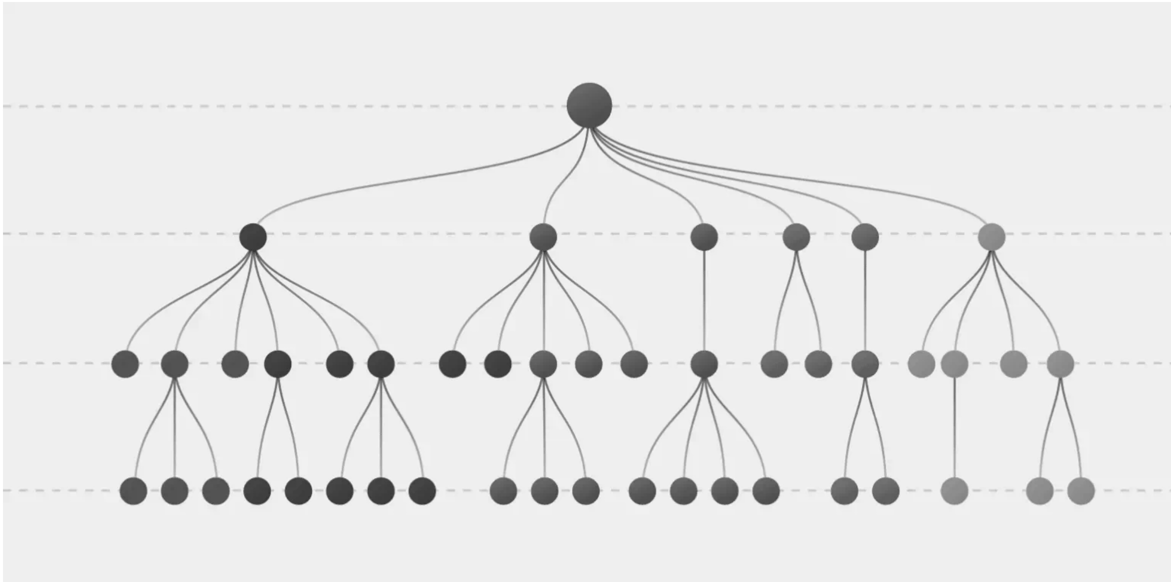

Figure 2.3: Decision Trees Structure

The creation of smaller decision trees necessitates the use of techniques such as a new node splitting measure. The node splitting measure is the most important of the approaches that may be used to build a decision tree. [Chandra and Varghese, 2009].

2.5.5 Random Forest

The random forest model is an excellent statistical learning tool. A random forest is a classifier that consists of a collection of tree-structured classifiers $h(x, k), k = 1, ...$, where the k are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x [Breiman, 2001].

The random forest, as the name suggests, is made up of a huge number of individual decision trees that work together as an ensemble. The random forest's various trees each spat out a class prediction, with the class with the highest votes becoming our model's prediction.

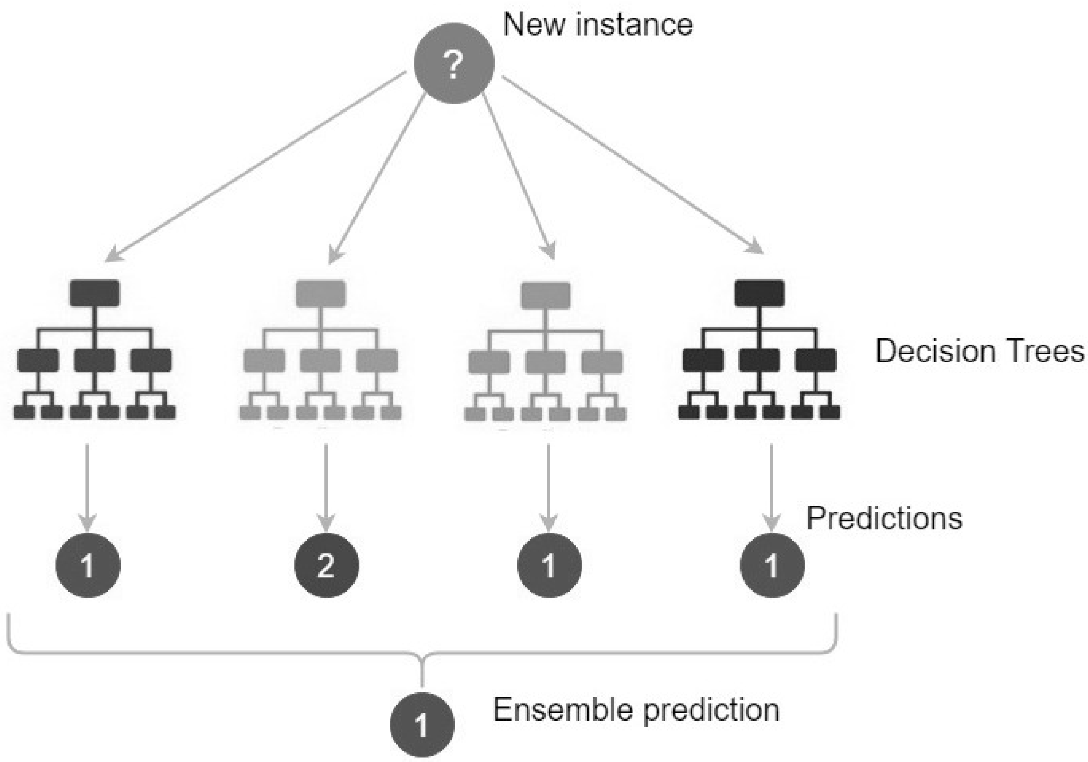

Figure 2.4: Random Forest Structure

2.5.6 Logistic Regression

Despite its name, logistic regression is a classification model rather than a regression model. For binary and linear classification problems, logistic regression is a simple and more efficient method. [Subasi, 2020].

According to [Caie et al., 2021] "In logistic regression, the conditional probability of the dependent variable (class) y is modeled as a logit-transformed multiple linear regression of the explanatory variables (input features) $x_1, ., x_n$ ":

$$P_{LR}(y = \pm 1|x, w) = \frac{1}{1 + e^{-yw^T x}} \tag{2.6}$$

2.5.7 Long Short-Term Memory

Long Short-Term Memory Networks (LSTM) are a type of recurrent neural network that may learn order dependence in sequence prediction problems. This is a requirement in a variety of complicated issue domains, including machine translation, speech recognition, and others.

The long-term dependency problem is explicitly avoided with LSTM. Remembering information for long periods of time is practically their default behavior, not something they struggle to learn!

2.6 Source codes examples

In this section, we will present some examples of source codes.

The instructions for loading the dataset are shown in figure 2.5

```
In [ ]: #Loading dataset

        import pandas as pd

        dataset = pd.read_csv('dataset.csv')
        dataset['text'].dropna(inplace=True)
        dataset['text'] = [entry.lower() for entry in dataset['text']]
```

Figure 2.5: Loading dataset

Instructions for displaying the dataset's head can be seen in Figure 2.6.

Figures 2.7 and 2.8 illustrate implementations of TfIdf and Word2Vec techniques for word embedding.

In [2]:
```
dataset.head()
```

Out[2]:

	text	sentiment
0	ما معني كل هذا ما معني كل هذا	0
1	من اسوا ما قرات ولا اجد حبكه او مغزي قمه الملل...	0
2	احلي تخلف	0
3	الله يرحم والديك الشيخ حفيظ علي هذا الكلام اكب...	0
4	زرت فرع الخبر المطعم شكله مستهلك واللحم ماله طعم	0

Figure 2.6: Dataset head

In [ ]:
```
#Tfidf vectorization

from sklearn.feature_extraction.text import TfidfVectorizer

Tfidf_vectorizer = TfidfVectorizer(min_df=5, max_df=0.95, sublin
ear_tf=True, use_idf=True, ngram_range=(1, 2))
Tfidf_vectorizer.fit(dataset['text'])
X_train_Tfidf = Tfidf_vectorizer.transform(X_train)
X_test_Tfidf = Tfidf_vectorizer.transform(X_test)
```

Figure 2.7: TfIdf vectorization

The instructions for calling the Classifiers NB, SVM, RF, DT, LR and LSTM are shown in Figures 2.9, 2.10, 2.11, 2.12 and 2.13, respectively.

```
In [ ]: import gensim from gensim.models
        import KeyedVectors from gensim.models
        import word2vec
        import numpy as np

        sentences = gensim.models.word2vec.LineSentence('corpus.txt')
        model = word2vec.Word2Vec(sentences, size=300,window=9,min_count
        =10,workers= 16)
        X = model[model.wv.vocab]
        words = list(model.wv.vocab)
        np.save('wordVectors.npy',X)
        np.save('wordsList.npy',words)
```

Figure 2.8: Word2Vec

```
In [ ]: #Naive Bayes Model
        from sklearn import naive_bayes

        NB = naive_bayes.MultinomialNB()
        NB.fit(X_train_Tfidf,y_train)
        predictions_NB = NB.predict(X_test_Tfidf)
```

Figure 2.9: The training code for the NB classifier

```
In [ ]: #SVM Model
        from sklearn import svm

        SVM = svm.SVC(C=1.0, kernel='linear', degree=3, gamma='auto')
        SVM.fit(X_train_Tfidf,y_train)
        predictions_SVM = SVM.predict(X_test_Tfidf)
```

Figure 2.10: The training code for the SVM classifier

```
In [ ]: # Random Forest Model
        from sklearn.ensemble import RandomForestClassifier

        RF = RandomForestClassifier(n_estimators = 10, criterion = 'entr
        opy', random_state = 0)
        RF.fit(X_train_Tfidf,y_train)
        predictions_RF = RF.predict(X_test_Tfidf)
```

Figure 2.11: The training code for the RF classifier

```
In [ ]: # Decision Tree Model
        from sklearn.tree import DecisionTreeClassifier

        DT = DecisionTreeClassifier(criterion = 'entropy', random_state
        = 0)
        DT.fit(X_train_Tfidf,y_train)
        predictions_DT = DT.predict(X_test_Tfidf)
```

Figure 2.12: The training code for the DT classifier

```
In [ ]: #Logistic Regression Model
        from sklearn.linear_model import LogisticRegression

        LR = LogisticRegression(random_state = 0)
        LR.fit(X_train_Tfidf,y_train)
        predictions_LR = LR.predict(X_test_Tfidf)
```

Figure 2.13: The training code for the LR classifier

2.7 Experiment and evaluation

The SA could be considered as a sentiment classification challenge from the perspective of machine learning (binary classification in our case). We employed four metrics to present the experimental results: accuracy, precision, recall, and F1-Score, which are calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{2.7}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.8}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.9}$$

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.10}$$

TP : True Positif

TN : True Negatif

FP : False Positif

FN : False Negatif

The results of these experiments are presented in Table 2.3. The SVM and LR classifiers outperform all other classifiers in all measures, as seen in this table.

Both classifiers have better performance. The LR model outperformed all other models, where the LR model reached 89%. The SVM model gave similar results to those obtained in the work of . The NB model and RF model are similar in accuracy. But with other measures, The NB model outperformed RF model. The DT model 75% as accuracy. The performance of both classifiers is better. The LR model outperformed all other models, with an 89% success rate. The SVM model produced results that were similar to those obtained by [Abdelli et al., 2019]. In terms of accuracy, the NB and RF models are comparable. However, the NB model outperformed the RF model on other metrics. The DT model has a 75% accuracy rate.

The LSTM model reach the 81% as accuracy.

Table 2.3: The obtained experimental results

Model	Accuracy	Precision	Recall	F1-Score
SVM	**88.29%**	83.17%	85.99%	85.66%
RF	86.28%	71.49%	79.94%	78.19%
NB	86.16%	78.73%	82.95%	82.29%
DT	76.31%	73.15%	75.07%	74.69%
LR	**89.52%**	81.98%	86.11%	85.58%
LSTM	81.22%	78.53%	76.31%	77.12%

2.8 Discussion

The Logistic Regression model had the best result of all trained models, 89.52%. In addition, the Support Vector Machines model scored 88.29%. The Naive Bayes, Random Forest, and Decision Trees models all produce positive results.

These results indicate that Logistic Regression and Support Vector Machines are better classifiers.

The results of the Long Short Term Memory experiment It's acceptable, but it could be better. Many factors influence these outcomes, including the size of the dataset and the quality of the word representation.

When we compare our work to other works that use other dialects, such as MSA, Tunisian, Jordanian, Algerian, and Saudi, we find that our work is superior. Our classifiers performed admirably, as we can see. The comparison is summarized in the table 2.4.

Despite accurate measurements, our model made some few errors. Table 2.5 contains a list of these errors.

The causes of these inaccuracies can be explained as follows:

- LABR and Multi domaine datasets are limited to some domain.
- LABR and Multi domaine datasets are written only in Arabic character, while the evaluation set contains Latin character.

Table 2.4: Comparison of the results of the various studies

Dialect	Classifier	Results	
		Accuracy	F-Score
Algerian (Our work)	**LR**	**89.52%**	**85.58%**
MSA [Ibrahim et al., 2015]	**SVM**	**90.62%**	**93.61%**
Tunisian [Medhaffar et al., 2017]	**MLP**	**/**	**78%**
Jordanian [Atoum and Nouman, 2019]	**SVM**	**73%**	**72%**
Algerian [Abdelli et al., 2019]	**SVM**	**86%**	**85%**
Saudi [Al-Twairesh et al., 2018]	**SVM**	**/**	**61.5%**

Table 2.5: Comparison of the results of the various studies

Example	Our annotation	Model result
عيد سعيد و مبارك و كل عام و أنت بألف خير إن شاء الله	1	1
انا حاب كاس رايب و ربع كسرة خير من لفريت	1	1
إذا عرفتو في اقل من خمسة ثواني دير جام	1	0
هذا الانسان ماهوش رجل وطني	0	0
مالقيتو ما ديرو جايبنا المصري هادا خماج	0	0

- The lexical differences between Algerian dialect, MSA and other dialects.

2.9 Conclusion

In this chapter, we examined sentiments on a dataset containing 49864 Algerian dialect comments labeled as follows: 24932 positive comments and 24932 negative comments. We used six different classifiers: Support Vector Machines (SVM), Decision Trees (DT), Random Forest (RF), Nave Bayes (NB), Logistic Regression (LR), and Long Short-Term Memory (LSTM). We compared the results of the six classifiers. We found that the Logistic Regression classifier achieves the correct Accuracy (89.52%). We provided examples of analytical errors in our model and explained how the model made them.

General Conclusion

General Conclusion

The objective of this study is to detect polarity in social network articles in two ways: a positive publication and a negative publication. The target of the project is to construct a Python application that uses a data source (Dataset format .CSV) that contains texts that have been annotated with values of 0 and 1. In order to categorize these texts.

We'll start by defining some of the terms used in this work. Following it, we concentrated on related works. Then we analyzed at sentiments in a dataset with 49864 Algerian dialect comments. Support Vector Machines, Decision Trees, Random Forest, Naive Bayes, Logistic Regression, and Long Short-Term Memory were among the six classifiers used. Finally, we presented our study's experimental results.

The results are highly encouraging, with an accuracy of 89.52 percent when using the Logistic Regression classifier.

For future work, we could cite:

- Use techniques like Glove [Pennington et al., 2014] and ELMo [Peters et al., 2018] for extraction of textual dataset.
- Other classification algorithms and features should be used.
- Based on three or five classes, sentiment analysis is performed.

Bibliography

[Abdelli et al., 2019] Abdelli, A., Guerrouf, F., Tibermacine, O., and Abdelli, B. (2019). Sentiment analysis of arabic algerian dialect using a supervised method. In *2019 International Conference on Intelligent Systems and Advanced Computing Sciences (ISACS)*, pages 1–6.

[Adouane and Dobnik, 2017] Adouane, W. and Dobnik, S. (2017). Identification of languages in algerian arabic multilingual documents. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 1–8.

[Al-Twairesh et al., 2017] Al-Twairesh, N., Al-Khalifa, H. S., Al-Salman, A., and Al-Ohali, Y. (2017). Arasenti-tweet: A corpus for arabic sentiment analysis of saudi tweets. *Procedia Computer Science*, 117:63–72.

[Al-Twairesh et al., 2018] Al-Twairesh, N., Al-Khalifa, H. S., Al-Salman, A., and Al-Ohali, Y. (2018). Sentiment analysis of arabic tweets: Feature engineering and a hybrid approach. *arXiv preprint arXiv:1805.08533*.

[Alhojely, 2016] Alhojely, S. (2016). Sentiment analysis and opinion mining: A survey. *International Journal of Computer Applications*, 150(6):22–25.

[Aly and Atiya, 2013] Aly, M. and Atiya, A. (2013). Labr: A large scale arabic book reviews dataset. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 494–498.

[Atoum and Nouman, 2019] Atoum, J. O. and Nouman, M. (2019). Sentiment analysis of arabic jordanian dialect tweets. *International Journal of Advanced Computer Science and Applications*, 10(2).

[Bayoudhi et al., 2015] Bayoudhi, A., Ghorbel, H., Koubaa, H., and Belguith, L. H. (2015). Sentiment classification at discourse segment level: Experiments on multi-domain arabic corpus. *J. Lang. Technol. Comput. Linguistics*, 30:1–24.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning archive*, 45(1):5–32.

[Caie et al., 2021] Caie, P. D., Dimitriou, N., and Arandjelović, O. (2021). Chapter 8 - precision medicine in digital pathology via image analysis and machine learning. In Cohen, S., editor, *Artificial Intelligence and Deep Learning in Pathology*, pages 149–173. Elsevier.

[Chandra and Varghese, 2009] Chandra, B. and Varghese, P. P. (2009). Moving towards efficient decision tree construction. *Information Sciences*, 179(8):1059–1069.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

[Duwairi et al., 2014] Duwairi, R. M., Marji, R., Sha'ban, N., and Rushaidat, S. (2014). Sentiment analysis in arabic tweets. In *Information and Communication Systems (ICICS), 2014 5th International Conference on*, pages 1–6.

[ElSahar and El-Beltagy, 2015] ElSahar, H. and El-Beltagy, S. R. (2015). Building large arabic multi-domain resources for sentiment analysis. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 23–34.

[Habash et al., 2012] Habash, N., Diab, M., and Rambow, O. (2012). Conventional orthography for dialectal arabic. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 711–718.

[Habash, 2010] Habash, N. Y. (2010). *Introduction to Arabic Natural Language Processing*.

[Ibrahim et al., 2015] Ibrahim, H. S., Abdou, S. M., and Gheith, M. (2015). Sentiment analysis for modern standard arabic and colloquial. *arXiv preprint arXiv:1505.03105*.

[Isabelle et al., 2019] Isabelle, G., Maharani, W., and Asror, I. (2019). Analysis on opinion mining using combining lexicon-based method and multinomial naïve bayes. In *Proceed-*

ings of the 2018 International Conference on Industrial Enterprise and System Engineering (IcoIESE 2018), pages 214–219.

[Liu, 2012] Liu, B. (2012). *Sentiment Analysis and Opinion Mining*.

[Malik and Mumtaz, 2019] Malik, A. S. and Mumtaz, W. (2019). Chapter 8 - electroencephalography-based diagnosis of depression. In Malik, A. S. and Mumtaz, W., editors, *EEG-Based Experiment Design for Major Depressive Disorder*, pages 171–203. Academic Press.

[Mataoui et al., 2016] Mataoui, M., Zelmati, O., and Boumechache, M. (2016). A proposed lexicon-based sentiment analysis approach for the vernacular algerian arabic. *Research on computing science*, 110(1):55–70.

[Medhaffar et al., 2017] Medhaffar, S., Bougares, F., Estève, Y., and Hadrich-Belguith, L. (2017). Sentiment analysis of tunisian dialects: Linguistic ressources and experiments. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 55–61.

[Mikolov et al., 2013] Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *ICLR (Workshop Poster)*.

[Pang and Lee, 2008] Pang, B. and Lee, L. (2008). *Opinion Mining and Sentiment Analysis*.

[Pang et al., 2002] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 79–86.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Édouard Duchesnay (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

[Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

[Peters et al., 2018] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 2227–2237.

[Saad and Ashour, 2010] Saad, M. K. and Ashour, W. (2010). Osac: Open source arabic corpora.

[Seerat and Azam, 2012] Seerat, B. and Azam, F. (2012). Opinion mining: Issues and challenges (a survey). *International Journal of Computer Applications*, 49(9):42–51.

[Subasi, 2020] Subasi, A. (2020). Chapter 3 - machine learning techniques. In Subasi, A., editor, *Practical Machine Learning for Data Analysis Using Python*, pages 91–202. Academic Press.

[Thakkar and Patel, 2015] Thakkar, H. and Patel, D. R. (2015). Approaches for sentiment analysis on twitter: A state-of-art study. *arXiv preprint arXiv:1512.01043*.

[Versteegh, 2014] Versteegh, K. (2014). *Arabic language*. Edinburgh University Press.

[Zaidan and Callison-Burch, 2014] Zaidan, O. F. and Callison-Burch, C. (2014). Arabic dialect identification. *Computational Linguistics*, 40(1):171–202.

[Zerrouki and Balla, 2017] Zerrouki, T. and Balla, A. (2017). Tashkeela: Novel corpus of arabic vocalized texts, data for auto-diacritization systems. *Data in Brief*, 11:147–151.

Sentiment Analysis of Algerian Dialect

1st Hamza Gattal
Department of Mathematics and Computer Science
University of Laarbi Tbessi
Tebessa, Algeria
201534039521@univ-tebessa.dz

2nd Dalila Bouras
LAMIS Laboratory and Department of Mathematics and Computer Science
University of Laarbi Tbessi
Tebessa, Algeria
dalilabouras1984@gmail.com

3rd Hakim Bendjenna
LAMIS Laboratory and Department of Mathematics and Computer Science
University of Laarbi Tbessi
Tebessa, Algeria
hbendjenna@gmail.com

***Abstract*—Sentiment analysis is a task of natural language processing which has recently attracted increasing attention. While significant work was directed toward the sentiment analysis of English text there is limited attention in literature toward the sentiment analytic of Arabic language and more especially Algerian dialect. This paper focuses on the various supervised Sentiment Analysis methods available in the existing literature. Further, this work presents the comparative study of different Sentiment Analysis algorithms on the basis of accuracy.**

***Index Terms*—Sentiment Analysis, Arabic Language, Algerian Dialect.**

I. Introduction

Day after day, the importance of Sentiment Analysis (SA) is getting bigger, especially in businesses and marketing, It can give an idea of what people like and don't like. The social networks platforms like Facebook and Twitter have a large number of users, and they share their views everyday on products, ideas, services, etc. Despite the large amount of comments and reviews, the sentiment analysis or opinion mining (OM) made it easier to access useful information. The sentiment analysis field is a classification of a decision of opinion as positive, negative or neutral. Its main purpose is to extract users opinions from social networks, for example, using automated techniques to define their positions on a topic, which is often expressed in text form.

The English language is receiving considerable interest from researchers in the field of sentiment analysis [Pang and Lee, 2008]. In contrast to the Arabic language, where there are a limited number of researches. Most of them focus on Modern Standard Arabic (MSA) [Duwairi et al., 2014], among which few number focus on Arabic dialects like Tunisian Dialect [Medhaffar et al., 2017], Saudi Dialect [Al-Twairesh et al., 2018], Jordanian Dialect [Atoum and Nouman, 2019], Algerian Dialect [Mataoui et al., 2016] and more.

The purpose of this work is to analyse sentiment in the Algerian dialect. We evaluated the [Abdelli et al., 2019], we applied several models that differ from the one used by them.

This paper is organized as follows: Section 2, describes the background of the method used to analyse sentiments. Section 3, present the Arabic language and the Algerian dialect. Section 4, proposed Arabic Algerian dialect Sentiment Analysis model. Section 5, provides the evaluation measures, the experimental results and the evaluation of this model. .

II. Background

Sentiment analysis has been practiced on a variety of topics. For instance, sentiment analysis studies for movie reviews, product reviews, and news and blogs [Thakkar and Patel, 2015].

The sentiment analysis approach's divided to three techniques [Alhojely, 2016], which are:

- Machine Learning based approach.
- Lexicon based approach.
- Hybrid/Combined approach.

In this paper, we use the Machine Learning based approach, we compared between five models; the Support-Vector Machine (SVM), Random Forest (RF), Naive Bayes (NB), Decision Tree (DT) and Logistic Regression (LR). All the above models cannot work directly with the text. Before using any model, there is a method called word embeddings.

A. Word Embeddings

Word embeddings or word representations or also word encoding, is a method to map words to vectors of real numbers.

The word embeddings has many methods like Word2vec and Term Frequency-Inverse Document Frequency (TF-IDF).

The Word2vec was created by [Mikolov et al., 2013], is a method represents each distinct word with a particular list of numbers called a vector, contain two models:

- The Continuous Bag-of-Words Model (CBOW).
- The Continuous Skip-gram Model.

The TF-IDF is is the calculation of two statistics: term frequency (TF) and inverse document frequency (IDF). There are many methods to calculate TF and IDF, and the below equations are the most used:

$$tf(t,d) = \sum_{t' \in d} f_{t',d} \tag{1}$$

$$idf(t,D) = logN/n_t \tag{2}$$

t' is the term t occurrence in document d, n_t is the number of documents contains t and N is the number of documents.

B. Support Vector Machines

Support Vector Machines (also support-vector networks) is a supervised machine learning model that uses classification algorithms for two-group classification problems. Developed at AT&T Bell Laboratories by [Cortes and Vapnik, 1995]. Mathematically, if the training data is linearly separable, then a pair (w, b) exists such as $w^T x_i + b \geq 1$ for all $x_i \in P$ and $w^T x_i + b \leq -1$ for all $x_i \in P$ with the decision rule given by $f_{w,b}(n) = sgn(w^T x_i + b)$, where w is termed as weighted vector and b as the bias. It is easy to show that when it is possible to linearly separate two classes an optimum separating hyperplane can be found by minimizing the squared norm of the separating hyperplane [Malik and Mumtaz, 2019]. The minimization can be set up as a convex quadratic programming (QP) problem: Minimize $\Phi(w) = \frac{1}{2}\|w\|^2$ subject to $y(wx + b) \geq 1$, $i = 1, 2, ..., l$ [Malik and Mumtaz, 2019].

C. Naive Bayes

Naïve Bayes algorithm is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. Naïve Bayes provides a mechanism for using the information in sample data to estimate the posterior probability $P(y \mid x)$ of each class y, given an object x. Naïve Bayes is based on:

$$P(y \mid x) = P(y)P(x \mid y)/P(x) \tag{3}$$

$$P(x \mid y) = \prod_{i=1}^{n} P(x_i \mid y) \tag{4}$$

$$P(x) = \prod_{i=1}^{k} P(c_i)P(x \mid c_i) \tag{5}$$

D. Random Forest

A random forest is a classifier consisting of a collection of tree-structured classifiers $h(x, k), k = 1, ...$ where the k are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x [Breiman, 2001].

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.

E. Decision Trees

Decision trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

The production of decision trees of smaller size requires a techniques such as a new node splitting measure. The node splitting measure is primary among the techniques that can be implemented during the construction of the decision tree [Chandra and Varghese, 2009].

F. Logistic Regression

Logistic regression, despite its name, is a classification model rather than regression model. Logistic regression is a simple and more efficient method for binary and linear classification problems [Subasi, 2020].

In logistic regression, the conditional probability of the dependent variable (class) is modeled as a logit-transformed multiple linear regression of the explanatory variables (input features) [Caie et al., 2021].

III. THE ARABIC LANGUAGE AND ALGERIAN DIALECT

The Arabic language is a Semitic language spoken in in a large area including North Africa, most of the Arabian Peninsula (Jazīrat Al-'Arab), and other parts of the Middle East.

Modern Standard Arabic is the official language of the Arab World. It is the primary language of the media and education.

The Arabic dialects are the true native language forms. They are generally restricted in use for informal daily communication.

The Modern Standard Arabic (MSA), or Modern Written Arabic, is the written form of the Arabic language, differs in a non-trivial fashion from the various spoken varieties of Arabic [Zaidan and Callison-Burch, 2014]. MSA is a term used mostly by Western linguists to refer to the variety of standardized, literary Arabic that developed in the Arab world in the late 19th and early 20th centuries. It is the official language used in academia, print and mass media, law and legislation, though it is generally not spoken as a mother tongue [Habash, 2010]. MSA is much more modern. MSA is primarily written not spoken.

The Arabic dialects also called colloquial Arabic, in contrast, are the true native language forms. They are generally restricted in use for informal daily communication.They are not taught in schools or even standardized although there is a rich popular dialect culture of folktales, songs, movies, and TV shows. Dialects are primarily spoken not written [Habash, 2010].

[Habash, 2010] and [Versteegh, 2014] give a breakdown for spoken Arabic dialects into groups:

- Egyptian Arabic: covers the dialects of Egypt and Sudan.
- Levantine Arabic: covers the dialects of Lebanon, Syria, Jordan and Palestine.
- Gulf Arabic: covers the dialects of Kuwait, United Arab Emirates,Saudi Arabia, Oman, Bahrain, and Qatar.
- North African (Maghrebi) covers the dialects of Morocco, Algeria, Libyan, Tunisia and Mauritania.
- Iraqi Arabic.
- Yemenite Arabic.

The Arabic script is used to write Arabic language, there are also many languages around the world like Persian, Kurdish are written using Arabic script. Arabic dialects are by default written in Arabic script although there are no standard dialectal spelling systems [Habash, 2010]. In the Arabic script two types of symboles for writting are: letters (Figure 1) and diacritics (Figure 2). writtes from right to left.

[Habash et al., 2012] define give a brief definition of letters and diacritics as follows: Arabic letters are written in cursive style in both print and script (handwriting). Diacritics are additional zero-width symbols that appear above or below the letters.

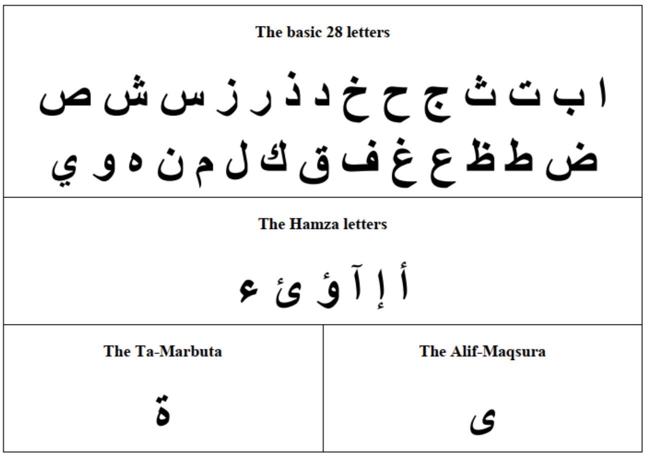

Fig. 1. The Arabic script letters

The Algerian dialect or Algerian Arabic is a group of North African Arabic dialects (Maghrebi) mixed with different languages spoken in Algeria [Adouane and Dobnik, 2017]. Algerian Arabic has a lot of influences from Berber as well as French, Classical Arabic, Modern Standard Arabic, and English [Habash, 2010].

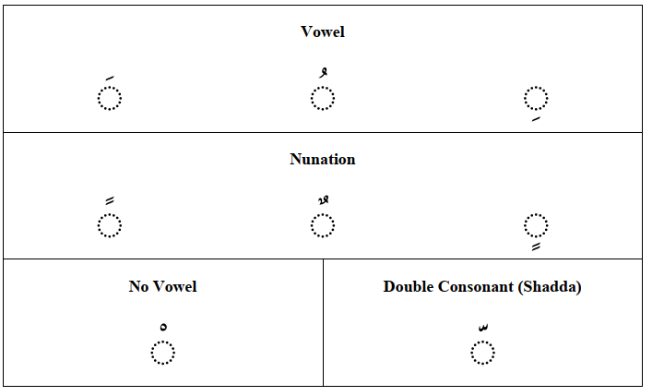

Fig. 2. The Arabic script diacritics

The Arabic as Modern Standard Arabic or Arabic dialects poses many challenges:

- A single dialect (for example the Algerian dialect) may contain several sub-dialects.
- There's a big difference between Modern Standard Arabic and Arabic dialects.
- A root of a word can take many forms depending on the context.
- The repetition of a letter several times to intensify the sense or feeling.
- The presence or absence of diacritics, can completely change the meaning of words.
- Words of negation used to deny verbs in the past or in the present, which change the meaning exactly to the contrary.

IV. Methodology

Basically, any Machine learning based analysis consists five steps:

- Data Collecting.
- Pre-processing.
- Training Data.
- Classification.
- Results.

A. *Data Collecting*

In our model, we use the [Abdelli et al., 2019] dataset. The dataset contains:

- The LABR [Aly and Atiya, 2013] dataset of book reviews.
- The multi domaine dataset of [ElSahar and El-Beltagy, 2015].
- The Algerian Arabic comments dataset of [Mataoui et al., 2016].
- 10k comments labeled into positive and negative, collected from popular Algerian Facebook pages.

The final dataset as shown in table I contains 49,864 comments that are evenly divided between positive and negative comments.

TABLE I
DATASET STATISTICS

	Positive	Negative
Total comments	24932	24932
Total words	1180663	1345029
Average words in each comment	47.36	53.95
Average characters in each comment	253.15	294.47

TABLE II
THE OBTAINED EXPERIMENTAL RESULTS

Model	Accuracy	Precision	Recall	F1-Score
SVM	88.29%	83.17%	85.99%	85.66%
RF	86.28%	71.49%	79.94%	78.19%
NB	86.16%	78.73%	82.95%	82.29%
DT	76.31%	73.15%	75.07%	74.69%
LR	89.52%	81.98%	86.11%	85.58%

B. Pre-processing

[Abdelli et al., 2019] mentioned that they have removed numbers and non-Arabic letters, and then normalized some letters as shown in figure 3.

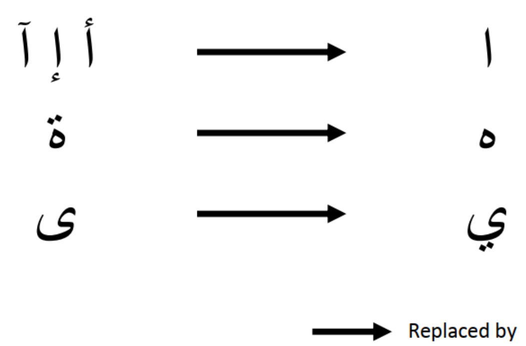

Fig. 3. Normalized letters

C. Training Data and Classification

To train five models, the dataset into a training set and test set, we gave 80% for training and 20% for test.

The models are:

- Support Vector Machines (SVM).
- Random Forest (RF).
- Naive Bayes (NB).
- Decision Trees (DT).
- Logistic Regression (LR).

For training all models, we used the scikit-learn (sklearn). For word embeddings, we used the TF-IDF techniques using the scikit-learn function TfidfVectorizer.

D. Results

To present The obtained experimental results, we used four metrics accuracy, precision, recall and F1-Score, which are calculated as follows:

- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)
- Accuracy = (TP + TN) / (TP + TN + FP + FN)
- F1 = 2 * (Precision * Recall) / (Precision + Recall)

TP (True Positif), TN (True Negatif), FP (False Positif), and FN (False Negatif).

The results obtained from conducting these experiments are shown in Table II. From this table, it is shown that the SVM and LR classifier performs better than other classifier in all measures.

Both classifiers have better performance on all measures. The LR model outperformed the SVM model, where the LR model reached 89% as accuracy and the SVM got just 88% as accuracy.

V. CONCLUSION

Sentiment analysis or opinion mining has increasingly evolved since the growth of social media networks; it is the process of evaluating the person's feelings to a specific subject.

The sentiment analysis models we have proposed in this paper is based on two classes/labels; positive and negative. These models are the SVM, RF, NB, DT and LR. Both of them are trained using the TF-IDF word embedding method. The dataset used for training and testing holds 49864 items. 80 % of the dataset is used for training and 20% is used for testing. Findings show that the LR Model outperform other models in term of accuracy.

This study is the baseline for our future work, where we plan to collect our dataset, based on three or five classes. also by appling various steps of preprocessing that includes; normalization, tokenization, name entity recognition, removing of stop word, and stemming. To get a very promising results.

REFERENCES

[Abdelli et al., 2019] Abdelli, A., Guerrouf, F., Tibermacine, O., and Abdelli, B. (2019). Sentiment analysis of arabic algerian dialect using a supervised method. In *2019 International Conference on Intelligent Systems and Advanced Computing Sciences (ISACS)*, pages 1–6.

[Adouane and Dobnik, 2017] Adouane, W. and Dobnik, S. (2017). Identification of languages in algerian arabic multilingual documents. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 1–8.

[Al-Twairesh et al., 2018] Al-Twairesh, N., Al-Khalifa, H. S., Al-Salman, A., and Al-Ohali, Y. (2018). Sentiment analysis of arabic tweets: Feature engineering and a hybrid approach. *arXiv preprint arXiv:1805.08533*.

[Alhojely, 2016] Alhojely, S. (2016). Sentiment analysis and opinion mining: A survey. *International Journal of Computer Applications*, 150(6):22–25.

[Aly and Atiya, 2013] Aly, M. and Atiya, A. (2013). Labr: A large scale arabic book reviews dataset. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 494–498.

[Atoum and Nouman, 2019] Atoum, J. O. and Nouman, M. (2019). Sentiment analysis of arabic jordanian dialect tweets. *International Journal of Advanced Computer Science and Applications*, 10(2).

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning archive*, 45(1):5–32.

[Caie et al., 2021] Caie, P. D., Dimitriou, N., and Arandjelović, O. (2021). Chapter 8 - precision medicine in digital pathology via image analysis and machine learning. In Cohen, S., editor, *Artificial Intelligence and Deep Learning in Pathology*, pages 149–173. Elsevier.

[Chandra and Varghese, 2009] Chandra, B. and Varghese, P. P. (2009). Moving towards efficient decision tree construction. *Information Sciences*, 179(8):1059–1069.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

[Duwairi et al., 2014] Duwairi, R. M., Marji, R., Sha'ban, N., and Rushaidat, S. (2014). Sentiment analysis in arabic tweets. In *Information and Communication Systems (ICICS), 2014 5th International Conference on*, pages 1–6.

[ElSahar and El-Beltagy, 2015] ElSahar, H. and El-Beltagy, S. R. (2015). Building large arabic multi-domain resources for sentiment analysis. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 23–34.

[Habash et al., 2012] Habash, N., Diab, M., and Rambow, O. (2012). Conventional orthography for dialectal arabic. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 711–718.

[Habash, 2010] Habash, N. Y. (2010). *Introduction to Arabic Natural Language Processing*.

[Malik and Mumtaz, 2019] Malik, A. S. and Mumtaz, W. (2019). Chapter 8 - electroencephalography-based diagnosis of depression. In Malik, A. S. and Mumtaz, W., editors, *EEG-Based Experiment Design for Major Depressive Disorder*, pages 171–203. Academic Press.

[Mataoui et al., 2016] Mataoui, M., Zelmati, O., and Boumechache, M. (2016). A proposed lexicon-based sentiment analysis approach for the vernacular algerian arabic. *Research on computing science*, 110(1):55–70.

[Medhaffar et al., 2017] Medhaffar, S., Bougares, F., Estève, Y., and Hadrich-Belguith, L. (2017). Sentiment analysis of tunisian dialects: Linguistic ressources and experiments. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 55–61.

[Mikolov et al., 2013] Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *ICLR (Workshop Poster)*.

[Pang and Lee, 2008] Pang, B. and Lee, L. (2008). *Opinion Mining and Sentiment Analysis*.

[Subasi, 2020] Subasi, A. (2020). Chapter 3 - machine learning techniques. In Subasi, A., editor, *Practical Machine Learning for Data Analysis Using Python*, pages 91–202. Academic Press.

[Thakkar and Patel, 2015] Thakkar, H. and Patel, D. R. (2015). Approaches for sentiment analysis on twitter: A state-of-art study. *arXiv preprint arXiv:1512.01043*.

[Versteegh, 2014] Versteegh, K. (2014). *Arabic language*. Edinburgh University Press.

[Zaidan and Callison-Burch, 2014] Zaidan, O. F. and Callison-Burch, C. (2014). Arabic dialect identification. *Computational Linguistics*, 40(1):171–202.

Manuscript Evaluation Form

Paper title	**Sentiment Analysis of Algerian Dialect**

Category	Mark with x
1. Original Scientific Paper	
2. Preliminary Communication	x
3. Review Article	
4. Expert paper	
Manuscript evaluation – Confirmation or Negating of qualification	Yes/No
1. Is the title clear and is it adequate to the purpose of the study	Yes
2. Abstract clearly presents objects, methods and results	Yes
3. Scientific methods are adequately used	Yes
4. Terminology is adequate	Yes
5. Paper does not have excesses and shortening is not necessary	No
6. Results are clearly presented	No
7. Conclusions are logically derived from the data presented	No
8. Key words are adequate	No
9. Supplements (tables, charts, pictures and drawings) are necessary and clear	Yes
10. References are appropriate	Yes

Reviewer comments
Through this paper, the author presented us with the various supervised methods of analysis of feelings available in the existing literature. Trying to also present a comparative study of different sentiment analysis algorithms based on accuracy. On the other hand, this study is the baseline for future work, hence the need to collect a data set, based on a number of classes. The author also had to show the appropriate steps for pre-treatment, standardization, name feature recognition, suppression of the stop word, and contain, in order to achieve more important results. I find that paper requires more development and deepening, especially in the comparison phase and the presentation phase of the results.

Decision	Mark with x
1. This manuscript is not acceptable	
2. This manuscript will be reconsidered after major revision	
3. This manuscript will be reconsidered after minor revision	x
4. This manuscript is acceptable in its present form	

Manuscript Evaluation Form

Paper title	Sentiment Analysis of Algerian Dialect

Category	Mark with x
1. Original Scientific Paper	
2. Preliminary Communication	x
3. Review Article	
4. Expert paper	
Manuscript evaluation – Confirmation or Negating of qualification	Yes/No
1. Is the title clear and is it adequate to the purpose of the study	No
2. Abstract clearly presents objects, methods and results	Yes
3. Scientific methods are adequately used	No
4. Terminology is adequate	yes
5. Paper does not have excesses and shortening is not necessary	
6. Results are clearly presented	No
7. Conclusions are logically derived from the data presented	Yes
8. Key words are adequate	yes
9. Supplements (tables, charts, pictures and drawings) are necessary and clear	Yes
10. References are appropriate	Yes

Reviewer comments
In this paper, the authors present a comparative study of different Sentiment Analysis algorithms on the basis of accuracy related to the supervised Sentiment Analysis methods for Algerian Dialect. 1. The title of the paper must be changed to be more significant. 2. Some language mistakes must be chicken 3. The database used must be presented with more detail. 4. In my opinion, the pre-treatment process seems incomplete. 5. The method used for the vectorial representation of the corpus to be used by a ML algorithm is not specified. 6. Table 1 represents the description of the database in which you mention only two classes (positive and negative) however, in the opinion mining process a third class (neutral) is used to classify the polarity of each opinion. 7. The choice of the dataset used is not justified. 8. What is the added value of your paper compared to the work of [Abdelli et al., 2019]. 9. In a research paper, and when we present a method or a process we must compare our work with similar contributions in the literature to mention the added value of our contribution. 10. this paper must be ameliorated to be accepted.

Decision	Mark with x
1. This manuscript is not acceptable	x
2. This manuscript will be reconsidered after major revision	
3. This manuscript will be reconsidered after minor revision	
4. This manuscript is acceptable in its present form	

Manuscript Evaluation Form

Paper title	Sentiment Analysis of Algerian Dialect

Category	Mark with x
1. Original Scientific Paper	
2. Preliminary Communication	X
3. Review Article	
4. Expert paper	
Manuscript evaluation – Confirmation or Negating of qualification	Yes/No
1. Is the title clear and is it adequate to the purpose of the study	Yes
2. Abstract clearly presents objects, methods and results	Yes
3. Scientific methods are adequately used	Yes
4. Terminology is adequate	Yes
5. Paper does not have excesses and shortening is not necessary	NO
6. Results are clearly presented	Yes
7. Conclusions are logically derived from the data presented	NO
8. Key words are adequate	Yes
9. Supplements (tables, charts, pictures and drawings) are necessary and clear	NO
10. References are appropriate	Yes

Reviewer comments
The purpose of this work is to analyse sentiment in the Algerian dialect using Machine learning apprch by investigating somme basic classifiers . The paper is well organized and claear but the idea behind isosimple and the contribution is not satisfying . Authors in this work use five (05) classic classfiers to test arabic dataset for sentiment analysis. Recently ,new techniques are proposed for automatic feature extraction (FE) of textual dataset like Glove , W2VEC and Elmo : Tf-idf method is too old and can't taking into account structural and semantic relations between words during FE propcess To improve this paper, Authors must incorporate one of those FE methods and must impove the discussion section. Conclusion section msut contain some futur purposes as perspectives

Decision	Mark with x
1. This manuscript is not acceptable	
2. This manuscript will be reconsidered after major revision	X
3. This manuscript will be reconsidered after minor revision	
4. This manuscript is acceptable in its present form	

Reviewer	
Name and surname	Nabiha AZIZI

To be returned by e-mail

e-mail: icrami2021@gmail.com

Thank you for taking the time and effort necessary to review the manuscript