SENTIMENT ANALYSIS CLASSIFICATION FOR TEXT IN SOCIAL MEDIA: APPLICATION TO TUNISIAN DIALECT

Asma BelHadj Braiek, Zouhour Neji Ben Salem

Department of Computer Sciences and quantitative methods
Faculty of Economic Sciences and Management of Nabeul
University campus, ElMrezgua, 8000
Carthage University, Tunisia

ABSTRACT

Social networks are the most used means to express oneself freely and give one's opinion about a subject, an event, or an object. These networks present rich content that could be subject today to sentiment analysis interest in many fields such as politics, social sciences, marketing, and economics. However, social network users express themselves using their dialect. Thus, to help decision-makers in the analysis of users' opinions, it is necessary to proceed to the sentimental analysis of this dialect. The paper subject deals with a hybrid model combining a lexicon-based approach with a modified and adapted version of a sentiment rule-based engine named VADER. The hybrid model is tested and evaluated using the Tunisian Arabic Dialect, it showed good performance reaching 85% classification.

KEYWORDS

automatic language processing, sentiment analysis, text mining, emotional detection, social web, annotated corpus, sentiment lexicon, and sentiment engine.

1. INTRODUCTION

Sentiment Analysis (SA) or Opinion Mining is one of the most popular information retrieval tasks for Natural Language Processing (NLP). SA identifies the presence of feelings or emotions expressed in a text, or a sentence. This emotion provides a positive, negative, or neutral feeling for a specific subject. Even though SA has progressed, the English language has been the most dominant language studied, and researches are more limited to other languages, including Arabic [24]. However, the Arabic language is taking an important part in communication via social networks, and the number of users using this language is increasing every day. This language is known for its richness and complexity and this makes the SA phase difficult, in addition to the lack of resources. Moreover, the language used in social networks is not Standard Arabic but the Arabic dialect, which makes the task of analysis even more difficult since the Arabic dialect differs from one country to another, and even within a country the dialect can vary from one region to another and inside a region, the spelling may differ from one person to another. Thus, the SA of the Arabic dialect, and especially the Tunisian dialect is a challenging task.

Tunisian Dialect is the primary form of Arabic spoken by Tunisian people in their daily lives and social texting, yet, it is still not well explored, and this can be explained by two main factors. First, the lack of additional resources like lexicon and corpus for this type of dialect. Second, the

David C. Wyld et al. (Eds): DBML, CITE, IOTBC, EDUPAN, NLPAI -2023
pp.313-326, 2023. IJCI – 2023 DOI:10.5121/ijci.2023.120223

complexity found in representation when dealing with it. The Tunisian dialect is very context-dependent and contains a variety of words that require different specific treatments [10].
When we deal with sentiment classification, there are essentially three approaches: Machine learning, Lexicon-based, and Hybrid technics [22]. Nevertheless, for Arabic SA, only a few papers use the hybrid method. A combination technic for Arabic SA was given by El-Halees [23]. The paper proposes a lexicon-based method combined with the Maximum Entropy approach (probabilistic classifier). This method has a classification accuracy of 80.29 %, according to the data. Another example of the hybrid model is HILATSA [25], which is a hybrid incremental learning technic for Arabic SA. The major aim of the paper is to develop a SA tool for Arabic tweets that can handle quick changes in words and their meanings. It combines the lexicon approach with machine learning models (SVM, LR, and RNN) to deal with varied word forms' and misspellings.The experiment demonstrated good results in a dynamic environment, with great precision and consistent performance.

The combination of the lexicon approach with a rule-based classifier is not well explored and we think it represents a third alternative of hybrid models that could perform well for SA. In fact, in our work, we will conduct a hybrid model because the hybrid approach includes the best features of machine learning and lexicon-based methodologies which allows the provisioning of a lexical representation of the Tunisian dialect with a good classification of the SA. The proposed hybrid model for the Tunisian dialect will combine a lexicon-based approach with an adapted rule-based approach by modifying a popular English tool "VADER" [18] to support the identification of the polarity of the Tunisian Arabic Dialect. Indeed, we modified the functionality of the English VADER tool, so that it can directly classify the sentiment of Tunisian texts without having to translate from Arabic to English.

Due to a lack of resources, we compiled a Tunisian polarity lexicon from an English polarity lexicon of SOCAL [20]. SOCAL is a dictionary of annotated words with their semantic orientation (polarity and strength).

To highlight our system, we have divided our work into three main parts. In section two, we focus on the particularities of the Tunisian Arabic Dialect and study works that had been done on SA. Section Threedescribes and explains in detail the design and the general architecture of the proposed hybrid model. Section four exposes and discusses the experimental results obtained. The paper is endedwith a conclusion summarizing the research process followed and presents some perspectives on future works.

2. TUNISIAN ARABIC DIALECT PARTICULARITIES AND RELATED WORKS ON SA

Tunisian Arabic Dialect (TAD) is a Maghrebi Arabic dialect spoken by more than 11 million people in their daily life. TAD is known as "Tounsi" /tu :nsi/ (which simply means Tunisian) or "Derja" (dialect). It is considered a low variety and an under-resourced language. It has neither a standard orthography nor dictionaries[1]. In addition, the friction of several languages throughout the history of the region has produced a complex and rich language comprising words, expressions, and linguistic structures. These languages are Berber, French, Italian, Spanish, and Turkish as well as other Romance languages Mediterranean [2]. TAD is nowadays very rich in lexical loans of multiple origins. It is also characterized by morphology, phonology, syntax, and a lexicon which have similarities and differences compared to MSA (Magrebien Standard Arabic) and even to other Arabic dialects [12].

2.1. Tunisian Arabic Dialect Spelling and Morphology

Like the Arabic language, TAD is written from right to left and each letter has an initial, medial, or final grapheme (shape) depending on its position in the word. TAD keeps the same short vowels as MSA,however,the vowel system of TAD is characterized, on one hand, by the transformation of long vowelsinto short vowels, especially when they are located in the final position of words, and on other hand, by the neglect of short vowels [3]. If a vowel is located at the end of a word carrying a single-syllable accent (e.g., jaA, /ja/, “he came”),),mšay, /mša/, he is party»)), it will be shortened.

TDA uses additional letters to the vowels and consonants of MSA, namely such as / g / ڨ/ and / v / ڥ. TAD is roughly composed of MSA, French (فرملي, (firmly) meaning ‘infirmier’ or ‘nurse’), and a mix of Turkish and Berber.

As TDA is an MSA derivative, words in TAD follow the exact representation of "root schema" as in MSA. In root schema, we can derive according to pre-established patterns by involving for example a vocalic variation or by adding certain elements from the MSA [3]. TAD adds some prefixes. For example, the root "خرج" translated into "He went out" is “ خْرَجْ” in TAD and none ” خَرَجَ” as for MSA. The two words have the same letter but different pronunciations (so they have different diacritics) leading to different schemas.

Inflectional affixes signal grammatical relationships, such as plural, past tense, and possession. They do not change the grammatical class of the stems to which they are attached [4]. As same as MSA, TAD has several inflectional morphological variations. We focused on distinguishing between the difference in moods, tenses as well as gender, and grammatical numbers. TAD has two types of Inflection:

1. ***In terms of Verb***: two main differences between TAD and MSA. The بْshows a comparison between verb morphological traits of the MSA and the TAD. Note that some dialects in Tunisia distinguish between singular male or female in the conjugation of the two verbs " write " (كَتَب) and "hit " (ضَرَبَ).It can be seen that the letters (ت,ن,ي) represent the conjugation prefixes and the letters (تي,بو, ت, بنا) represent suffixes for TAD. We also see that some verbs such as " تَكْتبْ" meaning "you write” is conjugated in the same way as in MSA. In other cases such as " نِكْتب" meaning, "I write" takes a different prefix which is (ن) instead of the prefix (أ) used in MSA for formingthe verb" أَكْتُبُ ". For the suffixes, they can also take another different one from MSA, for example, “تَضربُو ” meaning “you hit” that is conjugated in MSA “تضْرِبُون ” with « ون » as a suffix instead of « و » in TAD [2].
2. ***In terms of name***: the difference between TAD and MSA is that TAD is characterized by the absence of nominal marks. For this reason, it often loses the double nominal form which is extended by the numeral (زُوز, "two") followed by the plural. For example (استاذين, two professors) of MSA is translated as (زوزأساتذة, zuwzAasaAtdah,) for TAD [5].

- **Agglutination**: Like several Arabic dialects, affixes, and clitics in TAD have severaldifferences from MSA. TAD introduces new indefinite clitics that are not in MSA.In TAD, we distinguish two different ways to express the negation case. The first uses the words of negation (مغير,منغير,موش,مش), and the second uses the letters of negation ‘م,ش ‘either attached to the first position or the last position of the verb. It has the following form (ش + verb+ م, ش + verb+ ما).For example, the sentence " لن أتي اليوم للاجتماع " translated to “I will not come to the meeting” becomes ’ مش باش نجي ليوم للاجتماع’ in TAD using the case of the word ‘ مانيش باش نجي ليوم للاجتماع ‘ for the case of the letter. In contrast to MSA, in TAD the verb, the personal pronoun, and other pronouns are placed between these two negation pronouns.

- **Lexical:** The Tunisian lexicon has serval issues of many cultures, for this reason, the TAD vocabulary is much more different from the MSA. There are several words from French, Turkish, Italian, and Berber[5].

Since the years of 2010 and especially after the Tunisian revolution, some researchers focused on automatic TAD processing [2], [5], [6], [7]. Nevertheless, their work remains still preliminary. Some researchers have taken an interest in resource development (corpus, Wordnet, glossary, etc.) while others focused on the development of NLP tools. In this section, we will present an overview of the main works carried out for the benefit of TAD and SA. The TAD is a language that is very lacking in resources that can be directly exploited. As a result, creating a corpus for the TAD is a key step in automatic processing.

2.2. Related Works on Sentiment Corpus Construction

The creation of automatic language processing corpora is a necessary step in the treatment of a specific language. Karim Sayadi et al. [6] present a dataset collected by Twitter in the context of the Tunisian elections. They proposed a benchmark that can serve as a basis for future work. Another Tunisian Sentiment Analysis Corpus (TSAC) is presented by Medhaffar et al [7]. It has a total of 17,060 Tunisian comments received from Facebook. This corpus was compiled from comments left on the official pages of Tunisian radio and television stations.

Fsih et al [4] propose a statistical model for TAD to analyze and follow up on the opinions of Internet users upon broadcasts, TV, and radio programs. In [8], a lexicon-based SA system to classify the opinions embedded in Tunisian customers' reviews was proposed. To support identifying the Tunisian dialect, the author developed a Tunisian Arabic Analyzer along with a translating machine to handle the Arabizi format of the Tunisian dialect. Fourati et al[9] introduce TUNIZI a sentiment analysis Tunisian Arabizi Dataset, collected from Youtube social networks. The dataset, delivered by Tunisian speakers, is representative and preprocessed for analytical research.

Boujelbane[10] developed a bilingual MSA/TAD dictionary. She created it using a transformation method based on the grammatical categories of words in the corpus Arabic Treebank (ATB MSA) [11]. She creates a 500-word dictionary, using the pair (scheme, root) in MSA as a starting point, and then its TAD equivalent.

Abidi and Smaili[12] proposed a dictionary-based approach to analyze the sentiments of three Maghreb dialects, namely Algerian, Moroccan, and Tunisian. The dialect dictionary, used to classify documents by tone, is automatically created. These sentiment dictionaries were used to classify the three test datasets by polarity.

The TAD resources are not limited to the creation of corpora and lexicons, although two projects are underway to create a Wordnet for the TAD. The first is by Bouchlaghem et al. [13] where a Wordnet TUnDiaWN for TAD is under development using a corpus-based approach.

In [14], the author proposes a different approach to building a Wordnet for the TAD. The development of this Wordnet is based on two resources: the ANG/ TD "Peace Corps Lexicon" dictionary and the [15] corpus.

2.3. Works on Sentiment Analysis with a Machine-Learning Approach

Masmoody et al. [16] presented a work in the field of SA on two levels of Tunisian dialect, sentence level, and body level. They have selected reviews from the official Facebook pages of

Tunisian supermarkets, namely Aziza, Carrefour, Magasin General, Geant, and Monoprix. At the aspect level, it is the first work that addresses the problem of SA of Tunisian dialects. The SA task is done by a deep learning method. In fTheye developed three deep learning algorithms called CNN, LSTM, and Bi-LSTM. At the sentence level, their LSTM and Bi-LSTM systems achieve the best performance with an F-Measureof 87%. At the aspect level, the system achieves the best performance with an F-measure of 62% for the aspect category model and 78% for the sentimental model using Bi-LSTM and CNN.

3. TAD SA HYBRID MODEL: TUNISIAN VADER VERSION

In this part, we expose and implement a rule-based approach, From the lexicon of feelings, that determines the valence of a given message written in TAD. Our approach receives as input a message written in TAD (in Arabic letters) as well as TAD's pre-built sentiment lexicon. It returns message parity as output. Our approach consists of three main steps. Fig 1. shows the general structure of our model.

1. Corpus contraction
2. Lexicon construction approach
3. Rule-basedapproach

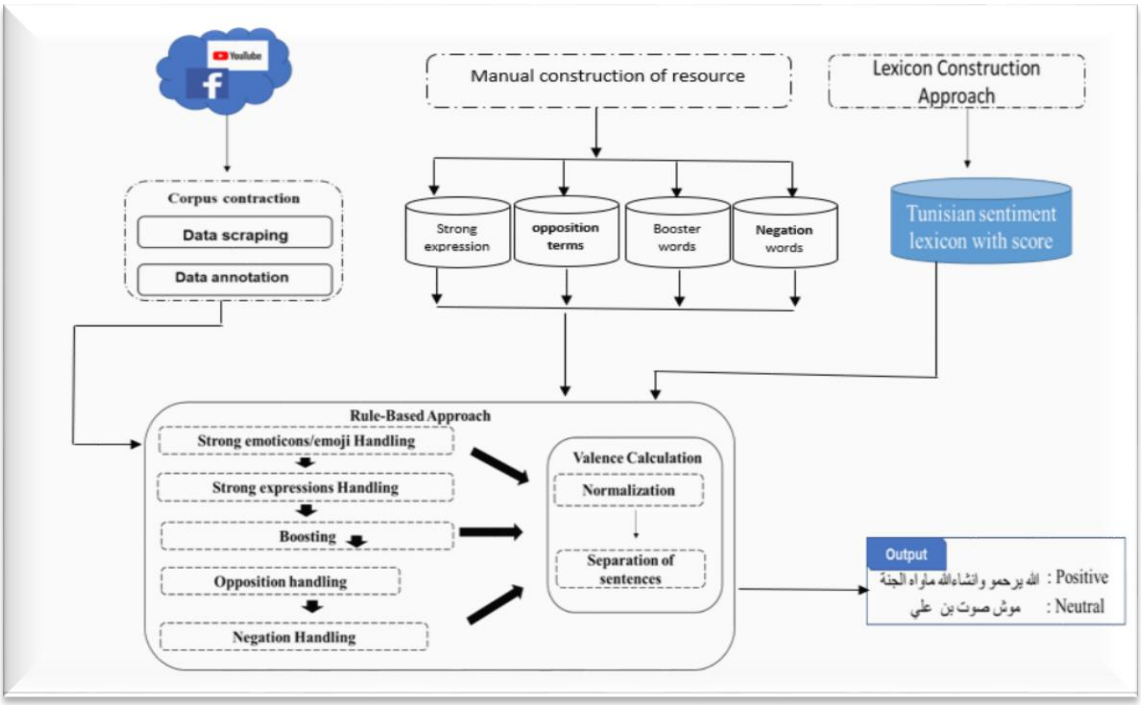


Figure 1. The general architecture of the Hybrid Model

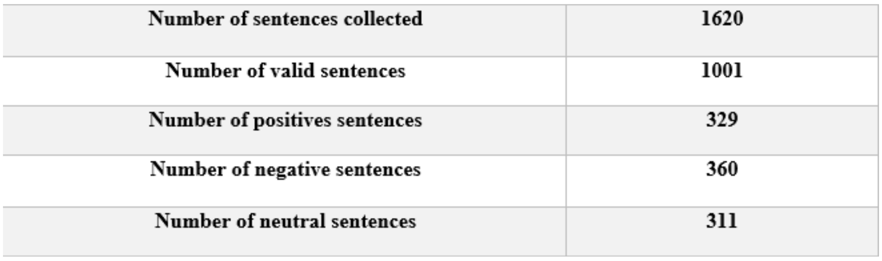
Data scrapped from Facebook and Youtube is preprocessed and annotated. In parallel, a scored Tunisian sentiment lexicon is generated by translating an English lexicon into Arabic. Rules from the Tunisian lexicon are manually defined and processed as well as the data and scored Tunisian lexicon into the rule-based Tunisian-Vadder Engine for SA classification. The three steps are more detailed in the following sections.

3.1. Corpus Contraction

The development of a powerful system for processing TAD requires the availability of a corpus. To build it, we decided to collect messages from social media platforms, since the cited corpus in the previous section is not for free. Indeed, until recently, dialectal Arabic did not have a written form, but social media has generated a large amount of dialectal data that can help us in the generation of the corpus.

The purpose of the corpus extraction step is to automatically extract data from social media. Facebook and YouTube are two of the most popular social media platforms in the world, especially in Tunisia. Therefore, we have used them to extract data answering our problem. For that, we determined, initially, a list including the identity of some Facebook as well as YouTube pages. Then, we extracted the comments from these pages in the appropriate dialect. For model efficiency, we proposed to collect a corpus in a suitable size. Then, we processed data scraping and finished the data annotation. Table 1 summarizes the data collection.

Table 1. constructed corpus statistics'



Number of sentences collected	**1620**
Number of valid sentences	**1001**
Number of positives sentences	**329**
Number of negative sentences	**360**
Number of neutral sentences	**311**

The sentences were labeled by two non-expert persons. The first step of double labeling was performed on 500 messages. It allowed us to detect a slight difference between the two annotators in the labeling task. Thus, we decided to compute the Inter-Annotator Agreements [21]. This metric is common practice in an annotation effort to compare annotations of a single source (text, audio, etc.) by multiple people. This is done for a variety of purposes, such as validating and improving annotation schemes and guidelines, identifying ambiguities or difficulties in the source, or assessing the range of valid interpretations (not to mention the study of annotation in its own right).

On the 500 double-labeled messages we have found (IAA=0.93). Since the IAA is high, we continued annotating the remaining 501 sentences.During Corpus Annotation, we faced some challenges and difficulties. We summarize them below and give some examples.

1. **Supplications:** We had difficulty determining the sentiment of supplications, as they can contain positive or negative words, but it is difficult to know if they express sentiment.
2. **Quotations:** These are in the form of inspirational quotes that generally convey a positive meaning, but do not constitute explicit advice about a specific target. We suggest that this type of text be considered neutral because it does not convey feelings.
3. **Determining the subject of the sentence:** it was sometimes difficult to determine the subject of some sentences and therefore we could not determine the sentiment expressed.

3.2. Lexicon Construction Approach

Our contribution is inspired by the work of Imane Guellil [17]. This work presents a tool for sentiment analysis of messages written in the Algerian dialect. Our TAD sentiment lexicon, shown in Fig 2, is created by extracting and scoring each sentence from the translated lexicon. Lexicon construction follows the three sub-steps:

- Translation of words from the English dictionary to MSA automatically using Google Translator.
- Manually Translation of words from the MSA into TAD. We tried to be as much exhaustive as possible in this step.
- Calculating the score of words in the new lexicon by summing the score of words that have the same meaning.
- A manual review step is needed to check Lexicon correctness because many words in the English language are found in the case of TAD.

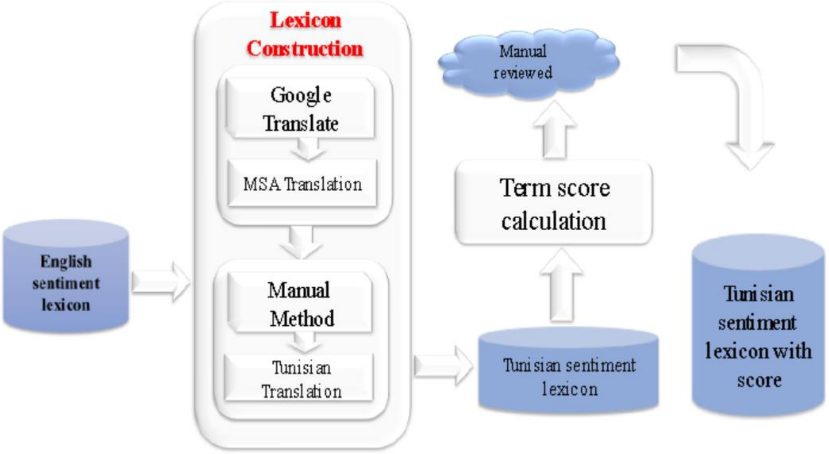
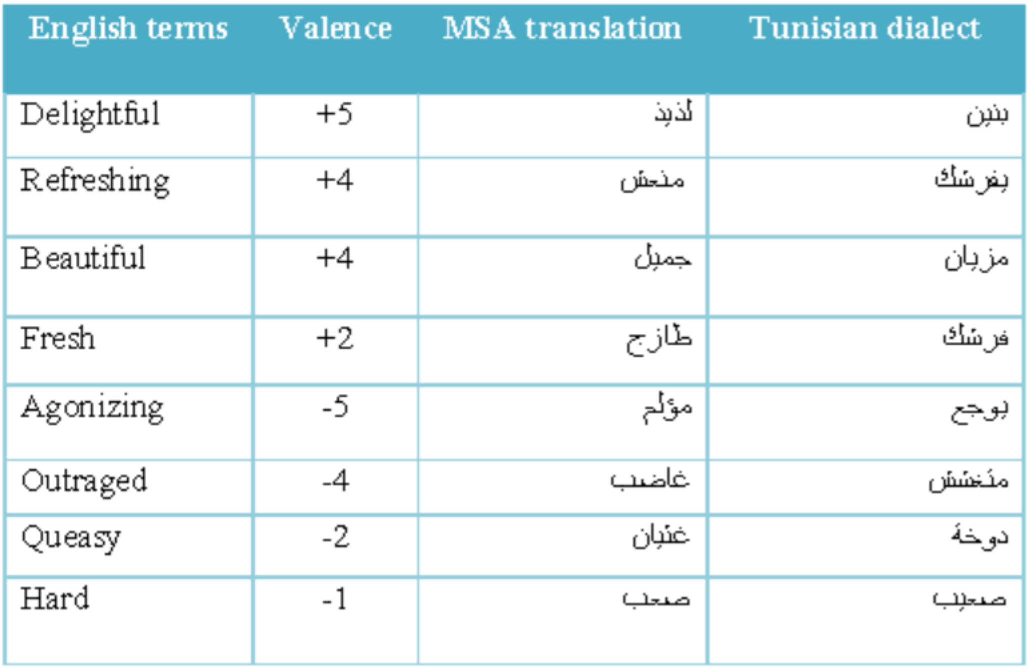


Figure 2. Lexicon construction approach

Table2 shows some examples of the translation from English, MSA, and then the Tunisian dialect.

Table 2. Examples of words translated in TAD with their valence



English terms	Valence	MSA translation	Tunisian dialect
Delightful	+5	لذيذ	بنين
Refreshing	+4	منعش	يفرشك
Beautiful	+4	جميل	مزيان
Fresh	+2	طازج	فرشك
Agonizing	-5	مؤلم	يوجع
Outraged	-4	غاضب	متغشش
Queasy	-2	غثيان	دوخة
Hard	-1	صعب	صعيب

3.3. Rule-Based Approach

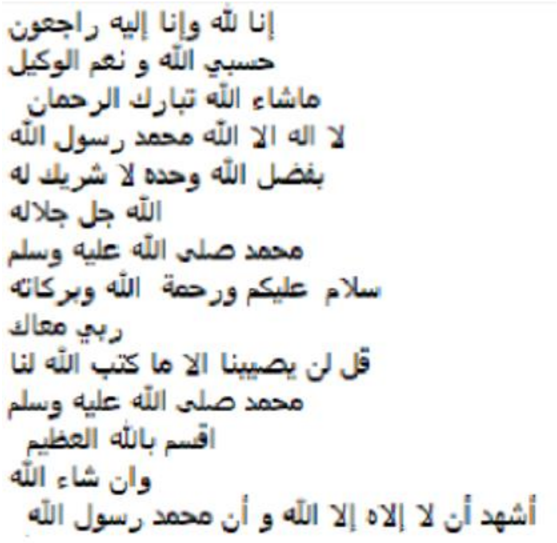
The goal of this step is to determine the polarity (positive or negative) of a message written in TAD as well as its score. This step is a classification problem requiring an annotated corpus to categorize the messages. In this step, our contribution is inspired by the work of an English Sentiment Analysis engine named VADER [18] [19].

As mentioned earlier, the rule-based approach is a supervised approach that belongs to the machine learning family. The work that first inspired our proposed rule-based method is that of Hutto and Gilbert [18] who presented one of the most famous sentiment analysis engines named VADER (Valence Aware Dictionary for sEntiment Reasoning) and available under the MIT license. VADER is based on three main steps:

- Development and validation of a reference sentiment lexicon that is sensitive to both the polarity and intensity of sentiments expressed in social media microblogs (but is also generally valid for sentiment analysis in other domains).
- Identification and experimental evaluation of generalizable rules regarding conventional uses of grammatical and syntactic aspects of the text to assess sentiment intensity.
- Comparison of the performance of a parsimonious lexicon and rule-based model with other established and/or typical sentiment analysis bases.

For our approach, we will define the main rules of the Tunisian dialect, inspired by VADER Heuristic rules, and take into account TAD characteristics in everyday life and social networks. In our work we will use the following five heuristics:
Strong emoticons and emoji rule: if a sentence contains sad emoji, that sentence is negative and if it contains a heart or positive emoji, it will be positive.

Strong expressions handling: The Tunisian people always express their feelings of joy and sadness through expressions that are specific to their culture. They can also express their wish to others using this type of expression. For example, if the message contains “الله يرحمو و ينعمو ” (meaning May God have mercy and bless) this message expresses a positive wish for a dead person. In contrast, if a message contains"حسبي الله و نعم الوكيل"then the message has a negative polarity. This rule is like the emoji rule, them eaning of the strong expression influences the classification of the message. A part of this list is shown in Fig 3.

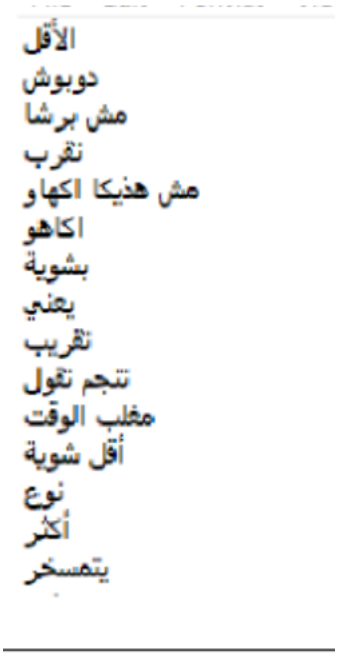


إنا لله وإنا إليه راجعون
حسبي الله و نعم الوكيل
ماشاء الله تبارك الرحمان
لا اله الا الله محمد رسول الله
بفضل الله وحده لا شريك له
الله جل جلاله
محمد صلى الله عليه وسلم
سلام عليكم ورحمة الله وبركاته
ربي معاك
قل لن يصيبنا الا ما كتب الله لنا
محمد صلى الله عليه وسلم
اقسم بالله العظيم
وان شاء الله
أشهد أن لا إلاه إلا الله و أن محمد رسول الله

Figure 3. Part of the list of positive and negative strong expression

Boosting words: Like English speakers, Tunisians use accent words to express the intensity of their feeling in messages. Around the world, the most commonly used accent word is " برشا" (which means "very, so much"). For example, if a Tunisian says " نحبك برشا" (meaning I love you so much), this word in the message means that the intensity of love is important.

This rule will be the same as for VADER, we look for whether the text contains a booster word existing in the dictionary of booster words. If it does, then it intensifies its valence according to the position of the booster word in the sentence. Next, we will check from the list of booster words, if it is a positive booster word Then the valence of the message will be incremented and if it is a negative booster, the valence will be decremented. Some booster words are presented in Fig 4.



الأقل
دوبوش
مش برشا
تقرب
مش هذيكا اكهاو
اكاهو
بشوية
يعني
تقريب
تنجم تقول
مغلب الوقت
أقل شوية
نوع
أكثر
يتمسخر

Figure 4. Example of booster words

Opposition handling: In MSA, and some dialects, such as the Tunisian dialect, the opposition is conveyed using the pronoun (meaning but, etc.). The valence is determined by the component after the opposition word, according to an analysis of a collection of communications including the opposition. The purpose of this rule is to search for the opposition words in the message. Afterward, the system focuses on the part after the opposition. The score is calculated by dividing the valence of the first part and then multiplying it by the valence of the second part.

Negation: Usually, the expression of feeling in a sentence can be affected by negative words. The use of negation words is different in the Tunisian dialect. In the case of English, it is used in the middle of sentences, but in the Tunisian dialect, it is used either at the beginning of the sentence, in the middle, or at the beginning of the words. At this stage, the negation word is searched according to the negation list built before. If a negation word is found, then the valence of the sentence is multiplied by -1. In TAD, a list of prefixes and suffixes related to negation is defined. It is composed of the following prefix and suffixes(مش(mush), موش(moush), مغير(magir), لا(lA), and منغير(mingyr)), ما(maA), م(mA) and ش(shA)).

Valence Calculation and normalization

The last step of the rules-based approach is to calculate the valence. This model guides in finding the correct valence for any text. At any time, the system outputs a sentiment score belonging to one of three classes: positive, negative, or neutral. To normalize the compound score, we add the valence of each word in the lexicon, adjusting it with rules, and then normalizing it to a range of -1 (extreme negative)to +1 (extremely positive) [18]. This is a handy metric for obtaining a single unidimensional sentiment assessment. It’s referred to as a "normalized weighted composite score." We apply the following equation to normalize the score:

Normalized score=(score)/(√score×score+alpha)

where alpha = 15 is the estimated maximum predicted value and score is the calculated score to be normalized. Negativity is expressed when the text's valence is less than 0and up to -1. If the valence is 0, the text is neutral, and if it is larger than 0 and up to +1, the text is positive. According to their determined polarity, each statement is labeled as positive, negative, or neutral [18].

4. Experimental Study

We present in this section all the data and parameters used in our experiments. We begin with the constructed corpus, then we describe the lexicon constructed, and finally test the corpora with the Tunisian Vader model. The final corpus consists of Tunisian comments that were collected between January 5 and February 24, 2022. We reached the collection of 1600 sentences and filtered them to obtain a total of 1001 valid sentences. Then, we start labeling the sentences according to polarity and language. The sentences were labeled by two non-linguist persons not experts in the domain.

Our sentiment lexicon is built by translating SOCAL (Semantic Orientation CALculator) into MSA and then into TAD. SOCAL is a dictionary of words annotated with their semantic orientation (polarity and strength) proposed by Taboada et al [20]. This lexicon is used to classify texts into positive and negative classes. The original version of SOCAL contains a total of 6,769 entries, including 1,142 verbs, 1,549 nouns, 2,827 adjectives, and 877 adverbs. We have chosen SOCAL for our model because it contains a large number of terms and allows us to give polarities independently of the context.

We apply our lexicon-based approach to TAD. As a baseline, we apply our approach to our corpus and the Naim-Mhedhbi corpus which is composed of those 8572 sentences (4229 Positive sentences, 1120 Negative sentences, and 3223 Neutral sentences). The study is conducted according to the above four measures mention in the three possible statement polarity: positive, negative, and neutral. Both include messages that have been manually annotated as positive and negative. In the context of our study, we use four metrics (Accuracy (%), Precision(P), Recall(R), and F1 Score (F1)) for evaluating the proposed rule-based SA approach as shown in Table 3.

Table 3. Experimental results and comparison with Naimi-Mhedhbi Corpus

	Positive Statement		**Negative Statement**		**Neutral Statement**		**Total Result**	
	Tunisian Vader corpus	Naim-Mhedh bcorpus	Tunisia n Vader corpus	Naim-Mhedhb corpus	Tunisia n Vader corpus	Naim-Mhedh corpus	Tunisian Vader corpus	Naim-Mhedhb corpus
Accuracy	0.848	0.92	0.846	0.953	0.854	0.8803	0.849	0.917
Precision	0.848	0.818	0.631	0.771	0.839	0.830	0.772	0.8064
Recall	0.739	0.942	0.793	0.856	0.754	0.77	0.762	0.856
F1 score	0.789	0.879	0.702	0.811	0.794	0.805	0.766	0.886

The results shown above are very encouraging as we reach a rate of 84,9% for accuracy and 77,2% for precision. Our Tunisian-Vadar engine outperforms the hybrid model presented in [23] combining the lexicon method with a probabilistic classifier and which reaches 82% of accuracy. It also gives better results compared to the HILATSA model which combines the lexicon approach with machine learning models and shows 73,67% accuracy for three classes of classification. The obtained result confirms what we have introduced that combining a lexicon

approach with a rule-based model could be a promising alternative to SA for the Arabic dialect. Therefore, we have tested our hybrid model with a more complete and richer corpus which is the one of Naim-Mhedhb [26].
From Table 3, Naim-Mhedhbi outperforms our constructed corpus, because of:

1. We have only a limited period for data collection while Naim-Mhedhbi has collected their corpus over a long period.
2. Naim-Mhedhbi is an expert in data science, so the selection and annotation of his corpus are one of the main factors that the result outperforms our results.
3. Table 3 shows also that our hybrid model is more performant in a larger corpus and can achieve excellent results reaching 91% of accuracy.

5. CONCLUSIONS AND FUTURE WORKS

In this work, we focused on Sentiment Analysis for the Tunisian Arabic Dialect. The latter is known to have low-resourced and to be a heterogenic language. We have created a sentiment analysis engine for the Tunisian Arabic dialect. The objective of this dissertation is corpus construction and polarity classification of the comments in the social networks for TAD messages.

For the realization of our purpose, we have been inspired by the work done on the lexical model VADER. We have also integrated some concepts used in the work of Guellil [17] for the sentimental analysis of the Algerian dialect. The study we conducted can be considered the first stone of TAD sentiment analysis with a rule-based approach and which brings a modification to a known classification system. TAD lexicon construction was one challenging aim that we have achieved and that could be improved later. The results we have obtained are very interesting and very promising and show that our Tunisian Vader Engine is very efficient and very competitive in the NLP field.

As perspectives, our system could be improved by considering the instability of the Tunisian dialect which lead to the change of meaning of certain words. We have also identified several research directions that deserve a more in-depth study. Therefore, improvements in the analysis of the Tunisian dialect are needed. In future work, we can extend the approach to cover the Arabizi classification.

REFERENCES

[1] E. Fsih, R. Boujelbane and L. H. Belguith, (2018) "Tunisian Dialect Resources for Opinion Analysis on Social Media," Joint International Conference on ICT in Education and Training, International Conference on Computing in Arabic, and International Conference on Geocomputing (JCCO: TICET-ICCA-GECO), pp. 1-7, doi: 10.1109/ICCA-TICET.2018.8726218.

[2] S.Mdhaffar, F.Bougares, Y. Estève, L.Hadrich-Belguith. (2017) "Sentiment Analysis of Tunisian Dialects: Linguistic Ressources and Experiments".Third Arabic Natural Language Processing Workshop (WANLP), Valence, Spain. pp.55-61.

[3] S. Mejri, M. Said, I.Sfar.(2009) "Pluringuisme et diglossie en Tunisie ', Synergies Tunisie n° 1 pp. 53-74

[4] D. Crystal, "A Dictionary of Linguistics and Phonetics". 6th ed., pp. 243-244. Malden, MA: Blackwell.

[5] I. Zribi, R. Boujelbane, A.Masmoudi, M.Ellouze, L.Belguith, and N.Habash. (2014)" A Conventional Orthography for Tunisian Arabic". In Proceedings of the Ninth International Conference on Language Resources and Evaluation, pages 2355–2361, Iceland.

[6] K. Sayadi, M. Liwicki, R. Ingold, and M. Bui, (2016)"Tunisian Dialect and Modern Standard Arabic Dataset for Sentiment Analysis: Tunisian Election Context," in Proceedings of the 17th International Conference on Intelligent Text Processing and Arabic Computational Linguistics, Konya, Turkey.

[7] S. Mdhaffar, F. Bougares, Y.Estève, and L. H. Belguith, (2017)"Sentiment Analysis of Tunisian Dialect: Linguistic Resources and Experiments," in Proceedings of the 3rd Arabic Natural Language Processing Workshop, pp. 55-61, Valencia, Spain.

[8] A. S. M.Alharbi, E. de Doncker. (2018) "Twitter sentiment analysis with a deep neural network: An enhanced approach using user behavioral information". Cognitive Systems Research, 54, 50-61.

[9] C.Fourati, A.Messaoudi, H. Haddad. (2020) " TUNIZI: A TunisianarabiZIsentiment analysis dataset ".

[10] R.Boujelbane. (2015) "Traitements linguistiques pour la reconnaissance automatique de la parole appliquée à la langue arabe : de l'arabe standard vers l'arabe dialectal". PhD Thesis, Université de Sfax et Aix-Marseille université.

[11] M. Maamouriand A.Bies. (2004)"Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools". In Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004).

[12] K. Abidi and K. Smaili. (2021)"Creating multi-scripts sentiment analysis lexicons for Algerian, Moroccan and Tunisian dialects". https://hal.science/hal-03308111/file/Papiersentiment_analysis_of_Maghrebi.pdf.

[13] R.Bouchlaghem, A.Elkhlifiand R.Faiz. (2014)"Tunisian dialect Wordnet creation and enrichment". In Arabic, Natural Language Processing Workshop co-located with EMNLP2014, Doha, Qatar.

[14] N.Karmani Ben moussaand M. Adel Alimi. (2015) "Construction d'un Wordnet standard pour l'Arabe Tunisien" .In CEC-TAL'2015.

[15] K. McNeil andF. Miled.(2011)"Tunisian Arabic Corpus: Creating a Written Corpus of an Unwritten Language". In Workshop on Arabic Corpus Linguistics (WACL), Lancaster University.

[16] A.Masmoudi, J. Hamdi and L.HadrichBelguith. (2021)"Deep Learning for Sentiment Analysis of Tunisian Dialect. Computación y Sistemas, Vol. 25, No. 1, 2021, pp. 129–148.

[17] I. Guellil, F.Azouaou, H. Saadane, and N. Semmar, (2017)"Une approche fondée sur les lexiques d'analyse de sentiments du dialecte algérien ". TRAITEME T AUTOMATIQUE DES LANGUES, 58(3), pp 41-65.

[18] C.J. Huttoand E. Gilbert.(2014)"VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text". Association for the Advancement of Artificial Intelligence.

[19] A. Amin, I. Hossain, A.Akther and K.MasudulAlam. (2019) "Bengali VADER: A Sentiment Analysis Approach Using Modified VADER". International Conference on Electrical, Computer and Communication Engineering (ECCE).

[20] M. Taboada, J. Brooke, M.Tofiloski, K.Voll, and M. Stede, (2011) " Lexicon-based methods for sentiment analysis". Computational linguistics, 37(2):267–307.

[21] R. Artstein, (2017)"Inter-annotator Agreement". In: Ide, N., Pustejovsky, J. (eds) Handbook of Linguistic Annotation. Springer, Dordrecht. pp 297–313.

[22] U. Kumari, D. Soni and A.K. Sharma, (2017) "A Cognitive Study of Sentiment Analysis Techniques and Tools: A Survey", IJCST vol.8, Issue 1.

[23] Alaa M. El-Halees, (2011) "Arabic opinion mining using combined classification approach",Proceeding The International Arab Conference.

[24] Guellil, I., Saâdane, H., Azouaou, F., Gueni, B., Nouvel, D., "Arabic Natural Language Processing: an overview", Journal of King Saud University - Computer and Information Sciences (2019), doi: https:// doi.org/10.1016/j.jksuci.2019.02.006.

[25] K. Elshakankery, M.F. Ahmed. "HILATSA: A hybrid Incremental learning approach for Arabic tweets sentiment analysis". Egyptian Informatics Journal. Volume 20, Issue 3, pp 163-171. 2019.

[26] https ://www.kaggle.com/datasets/naim99/tunisian-texts.

AUTHORS

Asma Belhaj Braiek is a Master of Research in Business computing student, she is preparing for her PhD in NLP and more precisely in the detection and classification of Tunisian Dialect Sentences that can offend the human sensation (sexual remarks, incitement to hatred, ...).

Zouhour Neji Ben Salem is an assistant professor in computer science. she conducts research in artificial intelligence and its various applications in the field of speech recognition, automatic processing of the Arabic language, and scheduling problems.