MATHEMATICAL MODELING AND COMPUTING, Vol. 11, No. 2, pp. 539–545 (2024)

Evaluating machine learning models efficacy in sentiment analysis for Moroccan Darija: An exploration with MAC dataset

Sakhi H., Elfilali S.

Faculty of Sciences Ben M'Sik, Hassan II University,
Bd Commandant Driss Al Harti, 7955, Casablanca, Morocco

(Received 18 December 2023; Accepted 25 June 2024)

Sentiment analysis is an essential technique for classifying and extracting emotions from several data sets. While many basic methods distinguish between negative and positive emotions, advanced approaches may consider additional categories, such as neutral emotions. This becomes very important and difficult when we need to deal with less parsed languages and dialects, such as Moroccan Darija. Our study highlights the nuances of conducting sentiment analysis implementing the MAC dataset, which includes comments in Moroccan Darija. Our main target is to do comparative study and research of the most used machine learning models for Arabic sentiment analysis, especially SVM, NB and KNN. These models have proven their effectiveness in classifying and analyzing emotions in widely studied languages such as English and Arabic. Through this comparative analysis, we aim to realize their effectiveness and adaptability in the Moroccan Darija dialect context.

Keywords: *ML; ASA; SA; opening mining; NLP; MAC.*

2010 MSC: 68T50, 68T01 **DOI:** 10.23939/mmc2024.02.539

1. Introduction

Sentiment analysis, is one of the important parts of text mining, it is dedicated to extract and interpret meaning from a variety of text sources, from survey responses to online reviews and comments on social media. At the heart of this task is the application of natural language processing (NLP) techniques, assigning sentiment scores such as -1 for negative sentiment and $+1$ for positive sentiment. In the digital age, every moment sees countless publications being shared online, especially on social platforms. This massive amount of data gives organizations the ability to monitor their brand or product awareness in real time [1].

Despite the enormity and ubiquity of this data, a significant amount of sentiment analysis research is still geared towards English, leaving other languages, especially variations of Arabic, to follow it [2,3]. Interestingly, comments in Arabic, especially in vernacular, have skyrocketed on the Internet. An emerging consensus indicates a growing need towards sentiment analysis in less studied languages, due to the lack of in-depth studies, especially in the Moroccan dialect.

Previous investigations have highlighted the effectiveness of various machine learning models, of which SVM emerged as a recurring model, showing a higher accuracy rate of over 82% in some studies [4]. These findings were corroborated by research focusing on perceptions of health services using the Twitter dataset. Here, models such as SVM, Naive Bayes and CNN have been evaluated, where SVM stands out for achieving an impressive accuracy of 91 [5].

In this context, our study introduces a new dimension by implementing MAC (Moroccan Arabic Corpus) [6] to analyze the emotions of comments written in the Moroccan dialect. Our contributions are threefold:

- describing and investigating the properties of the MAC dataset;
- presenting different techniques to preprocess the MAC dataset comments for sentiments analysis;
- presenting a set of experiments to the dataset to establish a comparison on different classifiers.

The rest of the paper is structured as below: in section 3, we present the extraction process used for the classification stage, the experimental results are presented in section 4 and conclude with the conclusion in section 5.

2. Statistics

The Arabic language, with its deep historical roots and vast influence, is among the five most spoken languages in the world [7]. However, in the specific field of sentiment analysis research, this often seems overshadowed. Representing 274 million native speakers, Arabic is an official tongue across a myriad of nations such as Egypt, Saudi Arabia, Algeria, Iraq, and Morocco, among others [8]. Collectively, these countries, along with others, contributed to the excess of 437 million Arabic speakers in 2017, an increase of 1.5% from the previous year. Figure 1 visually presents the hierarchy of the most spoken languages, placing Arabic right behind strong languages like English, Chinese, Hindi, Spanish, and slightly above with French.

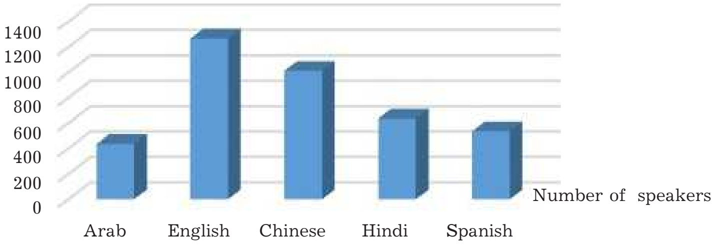

Fig. 1. Top 5 of the most spoken languages in the world.

But a stark contrast arises when the spotlight shifts to academic endeavors. While both Arabic and English sit atop global linguistic charts, the figure showcases the dearth of scientific contributions in NLP for the Arabic language. The English language sees a robust representation in sentiment analysis studies, whereas Arabic, despite its vast speaker base and official status in numerous countries, barely touches the 100-paper-per-year mark [6].

3. MAC (Moroccan Arabic Corpus) dataset description

The Moroccan Arabic Corporation (MAC) emerged as a pioneering force, presenting itself as the main and major open source dataset designed for sentiment analysis in the Moroccan dialect. It features an extensive collection of hand-tagged tweets that shed light on positive, negative or neutral emotions [9].

3.1. Properties of MAC

- Open source nature: MAC accessibility is one of its key features. Being open source ensures that researchers and developers can use its vast content for various analytical purposes.
- Dataset Composition: MAC has over 18 000 manually tagged tweets. These tweets are structured mainly in three areas: the content of the tweets, the content type and the sentiment class.
- Lexical Richness: The usefulness of the dataset is enhanced by its lexical dictionary, which includes 30 000 words. This vocabulary serves as a substantial repository, facilitating various NLP tasks.

3.2. Exploratory analysis

To truly harness the power of the MAC dataset, an exploratory analysis is essential. Digging deeper, we see that the dataset neatly classifies the comments into distinct classes. A quick look at the dataset reveals mostly positive responses.

3.3. Preprocessing steps

In the field of data science, preprocessing acts as a guardian, guarding the portals of accurate and insightful analysis. For MAC, this stage is very important to remove noise and refine the data set. The processing journey includes:

- delete duplicate tweets to avoid redundancy;
- filter all special characters, ensuring the purity of the text content;
- removing stop words often doesn't contribute to sentimental value;
- cut out repetitive characters for clarity and brevity;

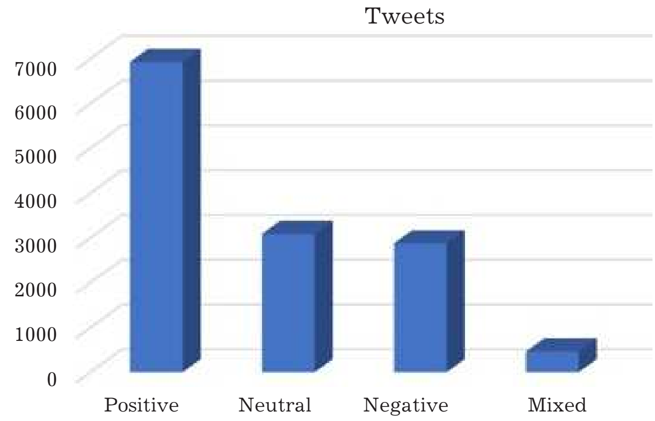

Fig. 2. MAC corpus statistics.

- refine hashtagged words by replacing "#word" with word", making them easy to understand;
- normalize data by removing diacritics, ensuring consistency;
- dataset coding — a step that converts phrases into lists of individual words.

3.4. Stemming and lemmatization

Both stemming and lemmatization serve as foundational pillars in the realm of natural language processing. These techniques are essential for distilling words to their basic nature, focusing mainly on extracting the base or root of a word. By doing so, these methods greatly reduce the complexity and size of the data set, while preserving its inherent semantic meaning.

Stemming: This process returns words to their original form by removing suffixes or prefixes. Finding the roots [10].

Lemmatization: Unlike root words, lemmatization converts words to their base form using language rules and dictionary look-ups. The key difference between root word (stemming) and Lemmatization is that lemmatization ensures that the root word (lemma) is a legal word in the language [11].

In the context of our research involving the MAC dataset, we chose ISRI Stemmer for this task. While stemming might seem rudimentary compared to lemmatization, the choice of ISRI Stemmer is well-founded. Designed to address the nuances and complexities of the Arabic language, this stemmer set excels at distilling words into meaningful and contextual roots. This ensures that despite the morphological richness of Arabic, our sentiment analysis models operate on a well-ordered dataset without losing significant semantic information [12].

3.5. Vectorization methods

Converting textual data to a format that machine learning algorithms can interpret is an essential step in natural language processing. This transformation, known as vectorization, represents the text as a digital vector, ensuring that the semantic nature of the content is maintained while making it processable by machine models. In our tests, we used two overall vectorization techniques:

- **TF/IDF (Term Frequency-Inverse Document Frequency):**
  - **Concept**: At its core, TF/IDF is all about weighing words. It assigns a weight to each word based on its frequency in the document relative to its frequency in the corpus [13].
  - **Significance**: Words that appear frequently in a document but are rare in many documents are often more informative. So, higher weights are assigned to them. This feature makes TF/IDF particularly adept at emphasizing words with contextual meaning, rather than generic terms that may appear all over the text [13].
  - **Application**: In the context of sentiment analysis, this can be very important because specific keywords can convey strong sentimental values. By highlighting these terms, TF/IDF can improve the accuracy of sentiment [13].
- **BOW (Bag of Words):**

- **Concept:** This method is relatively simple. BOW builds a vocabulary that includes all the unique words present in the data set. Each document (or text) is then introduced as a vector, representing the occurrence or frequency of each word in the vocabulary [14].
- **Significance:** Although simple, BOW is very powerful. It captures the essence of documents in terms of word frequency, making documents particularly well-suited to tasks where the occurrence of words is more important than their order or semantics [14].
- **Limitation:** The main limitation of BOW is its inability to capture contextual or word-order information, which can prove important in nuanced tasks [14].

4. Experimentation and results

4.1. Delving into the Core Algorithms for Arabic sentiment analysis

Sentiment analysis combines the complexities of linguistics and computing power. Therefore, choosing the right algorithm becomes paramount. Over the years, several algorithms have proven to be particularly adept at capturing the nuances of human emotions in text, especially when it comes to the complexity of the Arabic language. Let us take a deep dive into these chosen tools of the trade.

4.2. Support Vector Machines (SVM)

Conceptual framework. Basically, SVM embodies the idea that data, in multidimensional space, can be clearly separated by hyperplanes. SVM meticulously searches for the optimal hyperplane that gives the greatest distance between data layers [4],

$$y(x) = w \cdot x + b. \tag{1}$$

In the equation, $y(x)$ signifies the decision value for instance x, the sign of $y(x)$ is a determinant of the predicted class for $w \cdot x$ stands for the weight vector, while b is the bias.

Key Strengths. Very powerful in space characterized by many dimensions. Especially effective when the data has clearly separated margins [4].

Inherent Limitations. The computational demands of SVMs can be high, making them less suitable for huge data sets. They exhibit sensitivity to noise in the data and can be affected by outliers.

4.3. Nearest neighbor (KNN)

Conceptual framework: KNN operates on the principle of voting by majority of the neighborhood. Essentially, an unclassified data point is assigned the most common class among its data points, k nearest neighbors [15],

$$d(A,B) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}. \tag{2}$$

The above formula represents the Euclidean distance between two points A and B in n dimensional space. The closer the points are, more similar they are.

Key strengths. Due to its non-parametric nature, KNN makes no assumptions about the underlying data distribution. Its formula is simple, making it easy to implement and explain.

Inherent limitations. KNN is memory intensive because it requires storing the entire data set. The algorithm requires intensive computation, especially for data sets with a large number of data points.

4.4. Naive Bayes (NB)

Conceptual framework: Naive Bayes is a probability classifier, inspired by Bayes' theorem. He makes a bold assumption: predictors or characteristics that are independent of each other [16],

$$P(A/B) = \frac{P(B/A) \cdot P(A)}{P(B)}. \tag{3}$$

Here, $P(A/B)$ represents the posterior probability of (A) given (B), $P(B/A)$is the likehood, $P(A)$ is the class prior probability, $P(B)$ gives the predictor's prior probability [16].
Key Strengths: it is powerful in situations involving multi-class prediction. If the assumption of independence holds, New Brunswick can outperform many complex models.
Inherent Limitations: the basic assumption that the predictors are independent is strict. This is an idealization rarely seen in real-world datasets.

With these algorithms elucidated, we find ourselves armed with powerful tools, ready to dissect and analyze the intricate weave of sentiments in Arabic text, especially the Moroccan dialect. The goal is very clear: to test whether their proven effectiveness in the standard Arabic text is consistent with local dialectical nuances.

For each of our machine learning models: SVM, KNN, and NB we conducted dual experiments. One employed the TF/IDF vectorization while the other utilized the BOW method. The aim was to juxtapose the results and deduce the optimal vectorization strategy for sentiment analysis in the context of the Moroccan dialect.

In order to offer a comprehensive view of the comparative outcomes among SVM, KNN, and NB algorithms, we have visually represented the results in two formats.

Firstly, the capture bellow provides a graphical representation, making it easier to quickly grasp the performance trends and differentials of the algorithms.

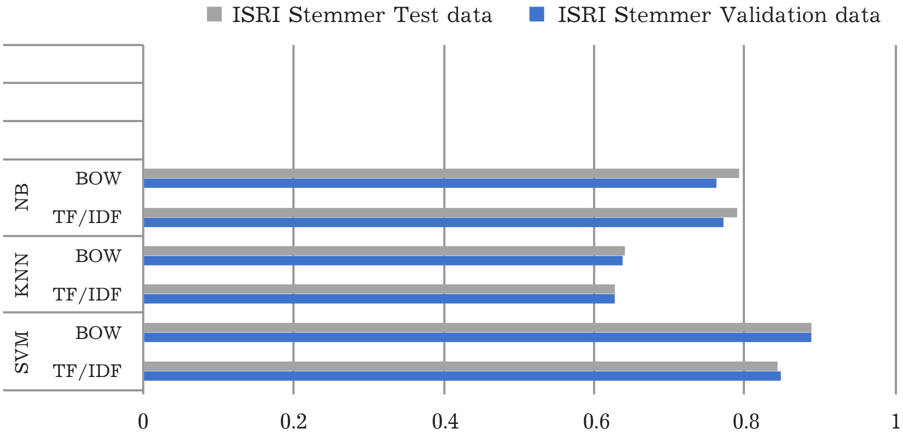

Fig. 3. Classification results in terms of accuracy.

Following that, for those desiring a detailed numerical breakdown, the following table tabulates the exact results. Each algorithm has been benchmarked using both vectorization methods while consistently leveraging the same stemming tool.

Table 1. Results and evaluations.

Stemming method: ISRI Stemmer			
Algorithms	**Vectorization methods**	**Validation data**	**Test data**
SVM	TF/IDF	0.846	0.843
	BOW	0.887	**0.889**
KNN	TF/IDF	0.628	0.627
	BOW	0.636	**0.639**
NB	TF/IDF	0.769	0.788
	BOW	0.760	**0.790**

The above table and accompanying diagram provide insights into the performance accuracy of various classifiers, keeping in mind that 70% of the corpus was set aside for validation and the remaining 30% for testing. Evidently, the SVM [4] outperforms its counterparts, showcasing superior results in comparison.

5. Conclusion

In this research, we harnessed the capabilities of the MAC database, a significant resource, to undertake an in-depth comparative assessment of the leading machine learning algorithms traditionally applied in sentiment analysis for Arabic. Our focal point in this study was comments expressed in the Moroccan dialect, an unique and intricate form of Arabic. Upon completing a meticulous preprocessing phase for our dataset, which involved cleaning, stemming, and other essential steps, we strategically divided it into two segments: a test portion comprising 30% and a larger validation segment making up 70%. This split was designed to provide a comprehensive evaluation of the algorithms' performance on both unseen and trained data. Our empirical results offer clear insights. SVM [4], among the evaluated algorithms, emerged as the standout performer, showcasing its robustness in handling the intricacies of the Moroccan dialect. It achieved a commendable prediction accuracy of 0.846 on the validation set and a closely aligned 0.843 on the test data. In comparison, while KNN [15] and NB [16] displayed commendable efforts, they fell short of the high benchmark set by SVM in this particular context.

[1] Liu B. Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. **5** (1), 1–167 (2012).

[2] Al-Ayyoub M., Khamaiseh A. A., Jararweh Y., Al-Kabi M. N. A comprehensive survey of arabic sentiment analysis. Information Processing and Management. **56** (2), 320–342 (2019).

[3] Keramatfar A., Amirkhani H. Bibliometrics of sentiment analysis literature. Journal of Information Science. **45** (1), 3–15 (2019).

[4] Ahmad M., Aftab S., Bashir M. S., Hameed N. Sentiment analysis using SVM: A systematic literature review. International Journal of Advanced Computer Science and Applications. **9** (2), 182–188 (2018).

[5] Alayba A. M., Palade V., England M., Iqbal R. Arabic language sentiment analysis on health services. 2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR). 114–118 (2017).

[6] Al-Ayyoub M., Khamaiseh A. A., Jararweh Y., Al-Kabi M. N. A comprehensive survey of arabic sentiment analysis. Information Processing & Management. **56** (2), 320–342 (2019).

[7] Arabic Fifth Most Spoken Language in the World: `www.Moroccoworldnews.com`.

[8] List of countries with Arabic as an official language: `www.wikipedia.org`.

[9] Garouani M., Kharroubi J. MAC: An Open and Free Moroccan Arabic Corpus for Sentiment Analysis. Innovations in Smart Cities Applications Volume 5. 849–858 (2022).

[10] Atwan J., Wedyan M., Bsoul Q., Hammadeen A., Alturki R. The Use of Stemming in the Arabic Text and Its Impact on the Accuracy of Classification. Scientific Programming. **2021**, 1367210 (2021).

[11] Freihat A. A., Abbas M., Bella G., Giunchiglia F. Towards an Optimal Solution to Lemmatization in Arabic. Procedia Computer Science. **142**, 132–140 (2018).

[12] Syarief M. G., Kurahman O. T., Huda A. F., Darmalaksana W. Improving Arabic Stemmer: ISRI Stemmer. 2019 IEEE 5th International Conference on Wireless and Telematics (ICWT). 1–4 (2019).

[13] Alammary A. S. Arabic Questions Classification Using Modified TF-IDF. IEEE Access. **9**, 95109–95122 (2021).

[14] Bekkali M., Lachkar A. Arabic Sentiment Analysis using Different Representation Models. International Journal of Emerging Trends in Engineering Research. **8** (7), 3368–3372 (2020).

[15] Duwairi R. M., Qarqaz I. Arabic sentiment analysis using supervised classification. 2014 International Conference on Future Internet of Things and Cloud. 579–583 (2014).

[16] Tan S., Cheng X., Wang Y., Xu H. Adapting Naive Bayes to Domain Adaptation for Sentiment Analysis. Advances in Information Retrieval. 337–349 (2009).

Оцінка ефективності моделей машинного навчання в аналізі тональності марокканської Дарії: дослідження з набором даних MAC

Сахі Х., Елфілалі С.

Факультет наук Бен М'Сік, Університет Хасана II,
бульвар Комендант Дрісс Аль Харті, 7955, Касабланка, Марокко

Аналіз тональності є важливою технікою для класифікації та вилучення емоцій із наборів даних сервера. Хоча багато базових методів розрізняють негативні та позитивні емоції, просунуті підходи можуть враховувати додаткові категорії, такі як нейтральні емоції. Це стає дуже важливим і складним, коли нам потрібно мати справу з менш аналізованими мовами та діалектами, такими як мароканська Дарія. Наше дослідження висвітлює нюанси проведення аналізу тональності за допомогою набору даних MAC, який включає коментарі марокканською мовою Дарія. Наша головна мета — провести порівняльні дослідження та дослідження моделей машинного навчання, які найчастіше використовуються для аналізу тональності на арабській мові, особливо SVM, NB та KNN. Ці моделі довели свою ефективність у класифікації та аналізі емоцій у таких широко вивчених мовах, як англійська та арабська. Завдяки цьому порівняльному аналізу намагаємося усвідомити їхню ефективність і адаптивність у контексті марокканського діалекту Дарія.

Ключові слова: *ML; ASA; SA; відкриття видобутку; NLP; MAC.*