НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
ноябрь–декабрь 2024 Том 24 № 6 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
November–December 2024 Vol. 24 No 6 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2024-24-6-982-990

Sentiment analysis of Arabic tweets using supervised machine learning

Ali Benabdallah[1]✉, Mohammed Alaeddine Abderrahim[2], Mohammed Mokri[3]

[1] Higher school of Management of Tlemcen, Tlemcen, 13000, Algeria
[2,3] University of Tlemcen, Tlemcen, 13000, Algeria

[1] benabdallah.a13@gmail.com✉, https://orcid.org/0009-0005-3312-7950
[2] abderrahim.alaa@yahoo.fr, https://orcid.org/0000-0001-7819-2354
[3] almakri.moh@gmail.com, https://orcid.org/0009-0008-6044-5235

Abstract
The increasing volume of user-generated content on social media platforms necessitates effective tools for understanding public sentiment. This study presents an approach to sentiment analysis of Arabic tweets using supervised machine learning techniques. We explored the performance of three popular algorithms — Support Vector Machines (SVM), Naive Bayes (NB), and Logistic Regression (LR) — on two distinct corpora: the Arabic Sentiment Text Corpus (ASTC) and a dataset of Arabic tweets. Our methodology involved four tests assessing the impact of corpus characteristics, preprocessing techniques, weighting methods, and the use of N-grams on classification accuracy. The first test established that the choice of corpus significantly influences model performance, with SVM showing superior accuracy on the structured ASTC, while NB excelled with the informal Arabic tweets. In the second test, preprocessing steps, including the removal of punctuation and stop-words, led to a noticeable improvement in classification accuracy for the Arabic tweets but had minimal or even negative effects on the ASTC. The third test indicated that incorporating N-grams yielded modest improvements for NB and LR in more structured texts, while its impact on tweets was negligible. Finally, the fourth test compared different weighting techniques, revealing that SVM benefitted from the Term Frequency-Inverse Document Frequency weighting method, while NB performance remained stable regardless of the weighting approach. These findings underscore the importance of tailoring preprocessing and feature extraction strategies to the specific characteristics of the dataset, ultimately enhancing the accuracy of sentiment analysis in Arabic language contexts.

Keywords
Arabic sentiment analysis (ASA), machine learning, classifier, polarity, Twitter

For citation: Benabdallah A., Abderrahim M.A., Mokri M. Sentiment analysis of Arabic tweets using supervised machine learning. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2024, vol. 24, no. 6, pp. 982–990. doi: 10.17586/2226-1494-2024-24-6-982-990

УДК 004.89

Анализ настроений арабских твитов с использованием контролируемого машинного обучения

Али Бенабдаллах[1]✉, Мохаммед Алаеддин Абдеррахим[2], Мохаммед Мокри[3]

[1] Высшая школа менеджмента Тлемсена, Тлемсен, 13000, Алжир
[2,3] Университет Тлемсена, Тлемсен, 13000, Алжир

[1] benabdallah.a13@gmail.com✉, https://orcid.org/0009-0005-3312-7950
[2] abderrahim.alaa@yahoo.fr, https://orcid.org/0000-0001-7819-2354
[3] almakri.moh@gmail.com, https://orcid.org/0009-0008-6044-5235

Аннотация
Растущий объем пользовательского контента на платформах социальных сетей требует эффективных инструментов для понимания общественных настроений. В работе представлен подход к анализу настроений арабских твитов с использованием контролируемых методов машинного обучения. Исследована производительность трех популярных алгоритмов — опорных векторных машин (Support Vector Machines, SVM), наивного байесовского алгоритма (Naive Bayes, NB) и логистической регрессии (Logistic Regression, LR) — на двух отдельных корпусах: арабском корпусе текстов настроений (Arabic Sentiment Text Corpus, ASTC)

и корпусе арабских твитов. Подход содержал четыре теста, оценивающих влияние характеристик корпуса: метода предварительной обработки, метода взвешивания и использования N-грамм на точность классификации. Первый тест позволил установить, что выбор корпуса значительно влияет на производительность модели, при этом SVM показал высокую точность на структурированном ASTC, в то время как NB лучше работает с неформальными арабскими твитами. Во втором тесте предварительная обработка, включая удаление знаков препинания и стоп-слов, привела к заметному улучшению точности классификации для арабских твитов, но оказала минимальное или даже отрицательное влияние на ASTC. Третий тест показал, что включение N-грамм дало незначительное улучшения для NB и LR в более структурированных текстах, в то время как его влияние на твиты было незначительным. Четвертый тест позволил сравнить различные методы взвешивания, показав, что SVM выиграл по сравнению с методом взвешивания TF-IDF, в то время как производительность NB оставалась стабильной независимо от подхода к взвешиванию. Полученные результаты подчеркивают важность адаптации стратегий предварительной обработки и извлечения признаков к конкретным характеристикам набора данных, что в итоге повышает точность анализа настроений в контекстах арабского языка.

Ключевые слова
анализ настроений на арабском языке (ASA), машинное обучение, классификатор, полярность, Twitter

Ссылка для цитирования: Бенабдаллах А., Абдеррахим М.А., Мокри М. Анализ настроений арабских твитов с использованием контролируемого машинного обучения // Научно-технический вестник информационных технологий, механики и оптики. 2024. Т. 24, № 6. С. 982–990 (на англ. яз.). doi: 10.17586/2226-1494-2024-24-6-982-990

Introduction

In recent years, there has been a rapid proliferation of social networking services, such as Facebook, Twitter, LinkedIn, Viadeo, Pinterest, etc. These social networks have enabled individuals and groups to express and share their opinions on various subjects (products, political events, economy, restaurants, books, hotels, video clips, etc.). Billions of comments and reviews are added every day on the web, making it necessary to explore user opinions to discover useful information. Manually exploring this enormous volume of comments and reviews is almost impossible. Thus, a new theme in Natural Language Processing (NLP) known as Sentiment Analysis (SA) or Opinion Mining has emerged [1].

The main objective of SA is to extract user sentiments/opinions from content created using automatic exploration techniques to determine their attitudes towards a subject often expressed in textual form. Nowadays, sentiment analysis is primarily used by companies to discover various customer opinions for marketing purposes [1–3]. It is also used in politics to predict election outcomes or to know public opinions on different policies. SA is also used in brand reputation management.

The field of SA is considered a classification task to determine if an opinion is positive, negative, or neutral [1] (and sometimes in other classes). SA approaches are based on one of the following classes: lexicon-based approaches, corpus-based approaches, and hybrid approaches [1, 4–7].

Most existing research on sentiment analysis focuses on English text [1, 3, 4, 8]. In recent years, researchers have tackled the challenges of sentiment analysis and opinion detection in Morphologically Rich Languages (MRLs). An MRL is a language in which important information about syntactic units and relations is expressed at the word level. Arabic is one of these languages that have begun to attract interest.

The goal of this work is to start a reflection to study SA in the case of the Arabic language. This document is organized as follows: the section “State of the Art” presents related work. The section “Our approach to Arabic sentiment analysis” describes our SA approach and presents our different datasets, and the section “Experimental Results” discusses the experimental results obtained. Finally, the section “Conclusion” provides the conclusion of our work.

State of the Art

In this section, related works on methods used for SA and opinion detection for the Arabic language will be discussed.

Al-Kabi et al. [9] developed a flexible and sizable standard Arabic Sentiment Analysis (ASA) corpus, intended as a foundational resource for constructing larger Arabic corpora. The corpus not only includes Modern Standard Arabic (MSA) but also reviews written in the five major Arabic dialects: Egyptian, Levantine, Arabian Peninsula, Mesopotamian, and Maghrebi. Additionally, it features five types of reviews: English, mixed MSA-English, French, mixed MSA-Emoticons, and mixed Egyptian-Emoticons. Released freely for researchers, this corpus is designed with flexibility, allowing users to modify its contents as needed. The initial version consists of 250 topics and 1,442 reviews, evenly distributed across five domains: Economy, Food and Lifestyle, Religion, Sports, and Technology, each containing 50 topics. The corpus was meticulously constructed manually to ensure high quality for researchers.

Oueslati et al. [10] noted the growing interest from the NLP research community and identified two primary approaches: the monolingual approach which relies on Arabic sentiment resources, and the bilingual approach which leverages English resources and machine translation. These studies provide a comprehensive overview of the current methodologies in ASA. The authors only covered articles published in the Springer, Elsevier, IEEE, ACM, and ACL databases.

Ghallab et al. [11] presented reviews and conferences on ASA indexed in Scopus, including several databases, such as Elsevier, Springer, and IGI Global. The authors provided a comprehensive review proposing taxonomy for sentiment classification methods. They highlighted the limitations of existing approaches, particularly in the preprocessing step, feature generation, and sentiment classification methods. Additionally, the study suggested potential trends for future research in ASA, both from practical and theoretical perspectives.

Rehab M. Duwairi and Raed Marji [12] applied SA on Arabic tweets to identify the polarity (positive, negative, or neutral) of the tweets. Their work involves testing the impact of stop word removal, negation detection, stemming, and converting words from dialect to standard Arabic (MSA) on the results of SVM, K-Nearest Neighbors (KNN), and the NB classifier.

Bolbol and Maghari [13] focus on sentiment analysis of Arabic tweets, conducting a performance comparison between three machine learning classifiers: LR, KNN, and Decision Tree. Using four Arabic text datasets, they evaluate the classifiers performance with four evaluation metrics: recall, precision, f-measure, and accuracy. The results indicate that LR achieves a better accuracy rate (93 %) on large datasets compared to the other classifiers.

Heikal et al. [14] used an ensemble model combining Convolutional Neural Network and Long Short-Term Memory models to predict the sentiment of Arabic tweets. Their model achieves an F1-score of 64.46 %, outperforming the state-of-the-art deep learning model F1-score of 53.6 % on the Arabic Sentiment Tweets Dataset.

Our approach to Arabic sentiment analysis

Challenges in Arabic Sentiment Analysis

Arabic is among the most widely spoken languages globally, particularly prevalent in the Arab world, especially throughout the Middle East and North Africa [15]. With 26 letters, Arabic is written from right to left and incorporates diacritical marks that aid in correct pronunciation and help distinguish between words with identical letters but differing meanings. There are three primary forms of the Arabic language: Classical Arabic which is used in religious and formal contexts; MSA, typically found in modern media [15]; and colloquial dialects which vary regionally across the Middle East and North Africa and lack standardization. According to [16], dialects dominate about 90 % of Saudi Twitter content compared to MSA. This poses challenges for researchers developing Arabic text classification models for SA. While translating MSA into English often yields good results, translating dialects can be difficult due to their heavy reliance on context [17, 18].

Used Arabic data source

In our experiments, we utilized two datasets: ASTC and Arabic tweets.

- The ASTC corpus (Arabic Sentiment Twitter Corpus): consists of four Tab-Separated Values files of 58,751 tweets (Table 1), with two columns each: one for the class and the other for the text of the tweets. Each line represents a tweet and its class: POS (Positive) or NEG (Negative) (Table 2).
- The second corpus we used is Arabic tweets:

It is composed of two ".txt" files, one for positive tweets and the other for negative tweets. (Table 3).

Each file contains 4 columns separated by tabs as follows:

Tweet_id, user_id, Tweet_Date, Tweet_text

The number of positive tweets is 10,000 and the number of negative tweets is 10,000.

Levels of SA

Research on sentiment analysis has primarily been conducted at three different levels of analysis: document level, sentence level, and aspect level [19].

- *Document Level:* The task at this level, known as document sentiment classification, aims to determine whether the overall opinion expressed in a document is positive or negative [19].
- *Sentence Level:* At this level, each sentence is considered an information unit that carries an opinion. The task at this level is to classify each sentence according to the opinion it expresses as a positive, negative, or neutral opinion. This level of analysis is considered a subjectivity classification, distinguishing between objective and subjective sentences (as illustrated in Fig. 1) [19].
- *Aspect Level:* Instead of looking at language constructs, the aspect level directly examines the opinion itself. It is based on the idea that an opinion consists of a sentiment and a target [19].

Table 1. Number of positive and negative tweets in the ASTC corpus

File type	POS	NEG	Total
Training file	23,879	23,121	47,000
Test file	5,970	5,781	11,751
Total	29,849	28,902	58,751

Table 2. The first 5 lines of the positive tweets file and the negative tweets file (ASTC)

0	POS	♥عليك الله قفل ميتينك...	0	NEG	☹ رحمه الله رحمه واسعه واسكنه فسيح جناته
1	POS	... طيار عراقي يسال برج المراقبة في سلطنة عمان عن…	1	NEG	راح فين كلامك الي كنتي بتقوليه ☹ رايكو في صوتي...
2	POS	□علمهم تكفي حسبي الله عليهم	2	NEG	... كلام أهلها انهم دخلوها مستشفى الامل لمدة شهر ...
3	POS	... وعن ذكر الله لا تغفلون لا إله إلا الله …	3	NEG	ليش عم تسبنا يا مخرف...
4	POS	... كثير من جمهور#الأهلي يقول لي عطانا التشكيلة #اله…	4	NEG	... السعادة_دايما_هي اشتقت ل حديثك ، وكأني لم احا...
…	…	...	…	...	...

Table 3. The first 5 lines of the positive and negative tweets file (Arabic tweets corpus)

Number	Tweet_id	User_id	Tweet Date	Tweet_text	Class
positive					
1	53969409731067905	@AliAlMullaa	1 Apr 11	الملا الساعة 3 الصبح @AliAlMullaa : @Jaber mmm	1
2	53969871058382848	@ALThaidyF	1 Apr 11	...الخرس..مافي شيء أكيد غير ال @moussa_a_alkhars	1
3	53969818331779072	@r5ton	1 Apr 11	☺هههههه من شق ثوب الناس شقو ثيابه @BOREDENKT	1
4	53969765043150848	@sunshinesud	1 Apr 11	..دي شوية صعبة. قريتها ب @esamaldeen @makavelli	1
5	53969733086740480	@YAlshatry	1 Apr 11	كيفججربي ☺ ترا عندي عينات هاهاهال @2233333333	1
…	…	...	…	...	...
negative					
1	1458463490	@Hazimov	5 Apr 09	): ... كان الله في عونك @Khalid	0
2	1458438484	@Khalid	5 Apr 09	... واختبار و.. ان case studyواخس شي لمن يكون عندك...	0
3	1456810178	@Aziz_MB	5 Apr 09	): يفتح Gmailشكرا سعودي نت لا المنسجر ولا...	0
4	1456597887	@Masrya	5 Apr 09	...حاجة تغيظ اوي لما بيبقى في حد في وشك طول النهار...	0
5	1456449209	@misfer	5 Apr 09	... طلعت اليوم حل ميجر الامس... بعض المرات الذكاء	0
…	…	...	…	...	...

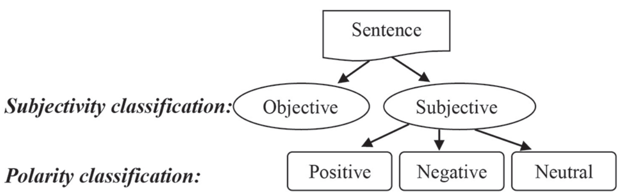

Fig. 1. SA at the sentence level

In our approach, we will work on the aspect level because it performs a more detailed and higher quality analysis, as it directly examines the opinion. Neither document analysis nor sentence analysis can precisely discover what people like and dislike.

Sentiment Analysis Process. There is a vast amount of existing work in the field of SA with researchers proposing various approaches. These can be summarized into three approaches [4]:
- Machine Learning-Based Approach;
- Lexicon-Based Approach;
- Hybrid Approach.

In our approach, we will focus on supervised machine learning systems. This involves two phases: training and testing, as illustrated in Fig. 2.

Regarding the classifiers used in our approach, we have opted for very popular algorithms which are: SVM, LR, and NB Classifiers.

Experimental Results

In this section, we present the experimental results obtained. We conducted experiments to highlight the

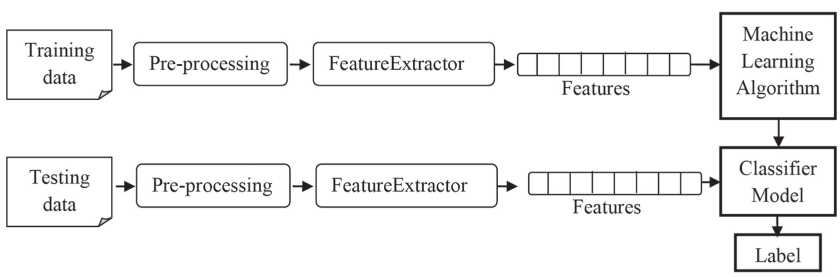

Fig. 2. Steps in our machine learning approach to sentiment analysis

importance and clarify the effect of some options on the performance of our sentiment analysis system.

We will conduct an initial experiment to test the impact of corpus construction (collection and labeling), a second to test the impact of corpus preprocessing and the classifier used, and a final experiment to see the effect of the choice of N-grams and weighting on the results. To achieve the most accurate results, we will use non-cross-validation first and cross-validation second.

Test 1. Testing the Impact of Corpus Construction (Collection and Labeling)

To do this, we adopted the following approach:

- Corpus: We used two corpora (ASTC and Arabic tweets).
- Preprocessing: No preprocessing.
- Feature Extraction and Representation:
  - Tokenization: Unigram;
  - Weighting: TF-IDF (Term Frequency-Inverse Document Frequency)
- Learning: We used three classifiers: SVM, NB, and LR.
  - The ASTC corpus is already divided into a training file and a test file;
  - For the Arabic tweets corpus, we divided it into two parts: 80 % for training and 20 % for testing.
- Testing and Measurement: After training the three classifiers (SVM, NB, and LR), we used precision, recall, F1-score, and accuracy as performance metrics. The test results based on accuracy are as follows.

According to the results shown in Table 4, we observe that:

- SVM provides good results with the ASTC corpus, slightly better than with the Arabic tweets corpus.
- NB yields good results with Arabic tweets but performs the weakest with the ASTC corpus.
- LR produces almost the same results, with a slight advantage in Arabic tweets.
- SVM and NB achieve the highest performance (80 %), the former with ASTC and the latter with Arabic tweets.

Linguistic Analysis

- Nature of the corpora: The ASTC corpus and the Arabic tweets corpus have significant differences that influence the results. The ASTC corpus is often more structured, typically used in contexts such as reviews or feedback (like customer reviews), where the language can be more formal and less ambiguous. In the Arabic tweets corpus, the language is more informal, including abbreviations, dialect expressions, and even code-switching with languages like English or French.

This linguistic difference between the two corpora partly explains why NB performs better on the second corpus (Arabic tweets). NB, due to its probabilistic approach based on word frequency, seems to better capture the nuances of informal language and common patterns on social media, where word distribution might be more regular.

- Dialect influence: Arabic tweets often include regional dialects, unlike the ASTC corpus, which might be dominated by MSA. This could explain why the performances differ between the two corpora. Models like SVM, which are more robust in formal contexts (such as with ASTC), may struggle more with dialectal variations, emojis, and linguistic ambiguity in Arabic tweets, where NB, being simpler, slightly outperforms SVM.
- Sentence length and structure: In ASTC, sentences are probably longer and grammatically more complex, whereas tweets are often short, concise, and sometimes lacking context. This can also influence the results of LR, which performs relatively evenly across the two corpora but with a slight improvement on Arabic tweets (0.774 vs 0.767). This could indicate that LR captures relationships between words well in short, structured sentences like tweets.

Computational Analysis

- SVM: It achieves high accuracy on both corpora, but performs slightly better on ASTC (0.800) than on Arabic tweets (0.790). This could be due to SVM ability to effectively separate data linearly when it's relatively structured, as in ASTC, where formal language makes this separation easier. On the other hand, the noisier nature of tweets may make it harder to distinguish between classes, leading to a slight performance drop.
- NB: This model performs better on Arabic tweets (0.800) than on ASTC (0.761). This can be explained by NB simplicity and efficiency on high-dimensional data like tweets, where words are often used repetitively or in predictable combinations. Tweets, with their concise and repetitive nature, may be better suited to a probabilistic approach, unlike ASTC, where the syntactic complexity reduces NB effectiveness.
- LR: This model performs similarly on both corpora, with a slight increase on Arabic tweets (0.774 vs 0.767). This suggests that LR is quite robust, even in the face of linguistic variations, and performs well with linear or near-linear data. However, the improvement on tweets could be due to the simpler and less structured nature of the sentences, making classification easier for LR.

In conclusion, these results show that the choice of corpus and linguistic characteristics significantly influence algorithm performance. SVM and LR are better suited to formal and well-structured corpora, while NB excels in more informal contexts where the model simplicity better captures lexical regularities.

Test 2. Evaluating the Impact of Corpus Preprocessing

To achieve this, we adopted the following approach:

- Corpus: We use two corpora (ASTC and Arabic tweets).
- Preprocessing: With preprocessing:
  - Removal of punctuation.
  - Removal of URLs.
  - Removal of @username.
  - Removal of HASHTAG # symbol.

Table 4. Accuracy results of the three classifiers for the two corpora (Test 1)

Classifier	SVM	NB	LR
ASTC corpus	0.800	0.761	0.767
Arabic tweets corpus	0.790	0.800	0.774

— Removal of stop-words: We used the NLTK list, which is a list of 750 words prepared by Mohataher Mohamed Alrefaie[1].
— Normalization of characters (أ،ي،ة،ك):
  — ("إ أ آ" to "ا")
  — ("ى" to "ي")
  — ("ة" to "ه")
  — ("ک" to "ك")
— Extraction and Presentation of Descriptors:
  — Tokenization: unigrams;
  Weighting: TF-IDF.
— Training and Testing: The same as Test 1, but we consider the remaining terms from the preprocessing step as descriptors.

After training the three classifiers SVM, NB, and LR, the test results in terms of accuracy are.

According to the results shown in Table 5, we observe that:

— For the Arabic tweets corpus, the performed preprocessing results in a significant improvement in the performance of all three classifiers.
— For the ASTC corpus, the preprocessing not only has no impact on the performance of the SVM and LR classifiers but also leads to a decrease in performance for the NB classifier.

The LR classifier with the Arabic tweets corpus achieved the best performance.

We can conclude that:

— Preprocessing can have a positive impact on the results of certain classifiers and for some corpora, but this is not always the case for other corpora.
— The removal of some text elements is not always beneficial because we might consider some texts or punctuation marks as noise and insignificant, but in reality these texts carry sentiment.

For example, an exclamation mark can indicate that the user is surprised by the very good quality of the product, or it can mean that the user is surprised by the poor quality of the product. Multiple exclamation marks "!!!!!!!!!!!!!" generally indicate a negative sentiment.

Linguistic Analysis

— ASTC corpus: The ASTC corpus is likely more formal and less prone to textual noise, such as hashtags, URLs, or mentions. This may explain why preprocessing had a negligible, or even negative, effect on performance for certain algorithms. The removal of stop-words and normalization of characters might have eliminated key information for NB, thus reducing its performance.
— Arabic tweets corpus: Tweets generally contain more noise, including contextual elements like hashtags, mentions, and frequent spelling variations. By removing these elements and normalizing characters, preprocessing allowed better isolation of sentiment cues, leading to improved performance across all models.

Computational Analysis

— SVM and LR: Both of these algorithms rely on linear separation of data and often benefit from cleaner and more coherent text. Their performance improved in the case of tweets, where preprocessing reduced noises, but remained stable with the ASTC corpus where the data was probably already sufficiently clean.
— NB: This simple, probability-based model relies heavily on word frequencies. In the ASTC corpus, preprocessing likely removed frequently used but informative words (like stop-words or specific word forms), whereas in tweets, it performed better due to noise reduction.

Table 5. Accuracy results of three classifiers for two corpora (Test 2)

Classifier	SVM	NB	LR
ASTC corpus	0.790	0.690	0.770
Arabic tweets corpus	0.850	0.860	0.880

In conclusion, preprocessing had varied effects depending on the nature of the corpus. For noisier texts, like tweets, it significantly improved algorithm performance, mainly by removing symbols and stop-words. However, for a more formal and structured corpus, like ASTC, the benefits of preprocessing were limited and could even reduce the performance of models like NB.

Test 3. Testing the Impact of N-grams

To do this, we adopted the following approach:

— Corpus: We use both corpora (ASTC and Arabic tweets);
— Preprocessing: No preprocessing.
— Extraction and Presentation of Descriptors:
  — Tokenization: Unigram, bigram, and trigram;
  — Weighting: TF-IDF.
— Training and Testing:
  — Same procedure as in Test 1.

After training the three classifiers (SVM, NB, and LR), the test results in terms of accuracy are as follows.

Based on the results shown in Table 6, we observe that:

— SVM. The results show that for both corpora the use of bigrams and trigrams did not improve performance compared to unigrams. The accuracy remains almost the same (0.800 for ASTC and 0.790 for Arabic tweets with unigrams), with a slight decrease observed when using bigrams and trigrams.

  This suggests that adding word relations for SVM through bigrams and trigrams do not provide much additional information. It is possible that unigrams already capture the essential features required for classification.
— NB. In the case of NB, there is a slight improvement in performance with the use of bigrams and trigrams for the ASTC corpus (from 0.761 to 0.771). For Arabic tweets, the results remain stable with a very slight decline when using trigrams.

  This can be explained by the fact that NB benefits slightly from capturing word relationships in the ASTC corpus, which may be more formal. However, for tweets, which are shorter and less structured, adding bigrams and trigrams might introduce too much noise.
— LR. It shows a gradual improvement with the addition of bigrams and trigrams for the ASTC corpus (from

[1] Mohataher Mohamed Alrefaie, Arabic-stop-words. Available at: https://github.com/mohataher/arabic-stop-words (accessed: 21.07.2024).

Table 6. Accuracy results of three classifiers for two corpora with (unigram, bigram, and trigram) (Test 3)

Classifier	SVM			NB			LR		
N-gram	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
ASTC corpus	0.800	0.800	0.799	0.761	0.769	0.771	0.767	0.776	0.778
Arabic tweets corpus	0.790	0.787	0.786	0.800	0.799	0.797	0.774	0.772	0.774

0.767 to 0.778). However, in the case of the Arabic tweets corpus, the results remain almost identical.

This suggests that LR is better able to capture lexical relationships in more structured texts like ASTC, where word pairs or triplets may add more meaning. For tweets, the informal and fragmented nature of the texts seems to limit the gains provided by N-grams.

Comparison between Algorithms

— NB vs. SVM and LR:

Unlike SVM and LR, NB shows slight improvement with bigrams and trigrams, especially in the ASTC corpus. This is consistent with NB nature, which is based on word or word pair probabilities and thus benefits from direct word relationships. However, the gains are modest.

SVM and LR, on the other hand, did not show significant improvements with bigrams and trigrams, which could indicate that these algorithms already find optimal decision boundaries using only unigrams.

Linguistic Analysis

— ASTC corpus: This corpus seems to benefit more from N-grams, especially for NB and LR. This suggests that word relationships are important in this corpus, perhaps due to *longer* and more complex sentences.

— Arabic tweets: Tweets, being generally *shorter* and more direct, do not show significant improvement with the addition of bigrams or trigrams. This is likely due to the fragmented nature of tweets, where lexical relationships between multiple words are less frequent or important.

In conclusion, adding lexical relationships through N-grams seems more useful in formal and structured corpora, while for shorter and more informal texts like tweets, unigrams remain sufficient to capture relevant information.

Test 4. Testing the Impact of Chosen Weighting

To conduct this test, we follow the following approach:

— Corpus: We used the Arabic tweets corpus.

— Preprocessing: Using preprocessing similar to that of Test 2.

— Feature Extraction and Presentation:

  — Tokenization: Uni-gram;

  — Weighting: CountVectorizer and TF-IDF.

— Training and Testing: Same procedure as in Test 1.

To compare the two weightings, we tested:

— The performance of three classifiers (SVM, NB, and LR) using CountVectorizer weighting once and TF-IDF weighting another time.

After training the classifiers, the classification results for Test 4 based on accuracy are shown in the following Table 7.

Based on the analysis of results shown in the Table 7 of Test 4, we observe that:

— TF-IDF weighting provides better results with SVM compared to CountVectorizer;

— Both weightings provide similar results with NB and LR.

Impact of Weighting Choice

— SVM: Switching from CountVectorizer to TF-IDF improved accuracy (from 0.7079 to 0.7358). This shows that SVM benefits from a more refined weighting method that considers the TF-IDF weighting, allowing more emphasis on distinctive terms.

— LR: The results for LR are very similar between CountVectorizer (0.7338) and TF-IDF (0.7326). This suggests that for this algorithm, the type of weighting has a limited impact. LR seems capable of exploiting both types of representation effectively, with no clear preference.

— NB: NB shows the same performance with both weightings (0.7426). This is expected since NB mainly works on probabilities based on term frequency, and it is less influenced by the complexity introduced by TF-IDF. CountVectorizer, which is based purely on word frequency, suits this algorithm well.

Computational Analysis

— SVM and the importance of TF-IDF: The fact that SVM performs better with TF-IDF indicates that this algorithm needs to capture the most important terms in the documents. TF-IDF reduces the weight of common words (such as stop-words) which helps SVM better separate the classes.

— NB Robustness: NB is relatively robust regarding the choice of weighting since it directly relies on term probabilities in documents. The similarity in results between CountVectorizer and TF-IDF shows that the adjustment of the inverse document frequency is not a determining factor for NB.

— LR and Independence from Weighting: LR shows very little difference between the two weightings. This might indicate that this algorithm is more flexible and can adapt to different data representations, whether through a simple weighting like CountVectorizer or a more complex one like TF-IDF.

Table 7. Classification results using CountVectorizer and TF-IDF weighting based on accuracy (Test 4)

Classifier	SVM	NB	LR
CountVectorizer	0.7079343399589734	0.7426956641048629	0.733816142587384
TF-IDF	0.7358314066333675	0.7426324261397578	0.732620021934390

Conclusion

In this study, we proposed an approach for sentiment analysis of Arabic tweets using supervised machine learning. Through four tests, we explored the impact of corpus construction, preprocessing, weighting techniques, and N-grams on the performance of three machine learning algorithms (SVM, Naive Bayes, and Logistic Regression) using two corpora (ASTC and Arabic tweets). The first test revealed that the choice of corpus and labeling plays a crucial role in the accuracy of the models. Our results further showed that preprocessing and the choice of weighting significantly affect SVM, while Naive Bayes remains stable regardless of the weighting method used. Additionally, the use of N-grams provided modest improvements for Naive Bayes and Logistic Regression in more structured corpora but had little impact on tweets. These findings highlight the importance of adapting preprocessing and feature extraction techniques to the specific characteristics of the corpus to improve classification accuracy.

References

1. Mataoui M., Zelmati O., Boumechache M. A Proposed lexicon-based sentiment analysis approach for the vernacular Algerian Arabic. *Research in Computing Science*, 2016, vol. 110, pp. 55–70. https://doi.org/10.13053/rcs-110-1-5
2. Al-Kabi M., Gigieh A., Alsmadi I., Wahsheh H., Haidar M. An opinion analysis tool for colloquial and standard Arabic. *Proc. of the fourth International Conference on Information and Communication Systems (ICICS 2013)*, 2013.
3. Pang B., Lee L. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2008, vol. 2(1-2), pp. 1–135. https://doi.org/10.1561/1500000011
4. Taboada M., Brooke J., Tofiloski M., Voll K., Stede M. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 2011, vol. 37, no. 2, pp. 267–307. https://doi.org/10.1162/coli_a_00049
5. Ding X., Liu B., Yu P.S. A holistic lexicon-based approach to opinion mining. *WSDM '08: Proc. of the 2008 International Conference on Web Search and Data Mining*, 2008, pp. 231–240. https://doi.org/10.1145/1341531.1341561
6. Kumar A., Sebastian T.M. Sentiment analysis on twitter. *IJCSI International Journal of Computer Science Issues*, 2012, vol. 9, no. 3, pp. 372–378.
7. Klenner M., Petrakis S., Fahrni A. Robust compositional polarity classification. *Proc. of the International Conference RANLP*, 2009, pp. 180–184.
8. Pak A., Paroubek P. Twitter as a corpus for sentiment analysis and opinion mining. *Proc. of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010.
9. Al-Kabi M., Al-Ayyoub M., Alsmadi I., Wahsheh H. A prototype for a standard Arabic sentiment analysis corpus. *International Arab Journal of Information Technology*, 2016, vol. 13, no. 1A, pp. 163–170.
10. Oueslati O., Cambria E., HajHmida M.B., Ounelli H. Review of sentiment analysis research in Arabic language. *Future Generation Computer Systems*, 2020, vol. 112, pp. 408–430. https://doi.org/10.1016/j.future.2020.05.034
11. Ghallab A., Mohsen Y., Ali Y. Arabic sentiment analysis: A systematic literature review. *Applied Computational Intelligence and Soft Computing*, 2020, vol. 2020, pp. 403128. https://doi.org/10.1155/2020/7403128
12. Duwairi R., Marji R., Sha'ban N., Rushaidat S. Sentiment Analysis in Arabic tweets. *Proc. of the 2014 5th International Conference on Information and Communication Systems (ICICS)*, 2014, pp. 1–6. https://doi.org/10.1109/iacs.2014.6841964
13. Bolbol N.K., Maghari A.Y. Sentiment analysis of Arabic tweets using supervised machine learning. *Proc. of the 2020 International Conference on Promising Electronic Technologies (ICPET)*, 2020, pp. 89–93. https://doi.org/10.1109/ICPET51420.2020.00025
14. Heikal M., Torki M., El-Makky N. Sentiment analysis of Arabic Tweets using deep learning. *Procedia Computer Science*, 2018, vol. 142, pp. 114–122. https://doi.org/10.1016/j.procs.2018.10.466
15. Alhamid M., Alsahli S., Rawashdeh M., Alrashoud M. Detection and visualization of Arabic emotions on social emotion map. *Proc. of the International Symposium on Multimedia (ISM)*, 2017, pp. 378–381. https://doi.org/10.1109/ISM.2017.76
16. Al-Thubaity A., Alqahtani Q., Aljandal A. Sentiment lexicon for sentiment analysis of Saudi dialect tweets. *Procedia Computer Science*, 2018, vol. 142, pp. 301–307. https://doi.org/10.1016/j.procs.2018.10.494

Литература

1. Mataoui M., Zelmati O., Boumechache M. A Proposed lexicon-based sentiment analysis approach for the vernacular Algerian Arabic // Research in Computing Science. 2016. V. 110. P. 55–70. https://doi.org/10.13053/rcs-110-1-5
2. Al-Kabi M., Gigieh A., Alsmadi I., Wahsheh H., Haidar M. An opinion analysis tool for colloquial and standard Arabic // Proc. of the fourth International Conference on Information and Communication Systems (ICICS 2013). 2013.
3. Pang B., Lee L. Opinion mining and sentiment analysis // Foundations and Trends in Information Retrieval. 2008. V. 2(1-2). P. 1–135. https://doi.org/10.1561/1500000011
4. Taboada M., Brooke J., Tofiloski M., Voll K., Stede M. Lexicon-based methods for sentiment analysis // Computational Linguistics. 2011. V. 37. N 2. P. 267–307. https://doi.org/10.1162/coli_a_00049
5. Ding X., Liu B., Yu P.S. A holistic lexicon-based approach to opinion mining // WSDM '08: Proc. of the 2008 International Conference on Web Search and Data Mining. 2008. P. 231–240. https://doi.org/10.1145/1341531.1341561
6. Kumar A., Sebastian T.M. Sentiment analysis on twitter // IJCSI International Journal of Computer Science Issues. 2012. V. 9. N 3. P. 372–378.
7. Klenner M., Petrakis S., Fahrni A. Robust compositional polarity classification // Proc. of the International Conference RANLP. 2009. P. 180–184.
8. Pak A., Paroubek P. Twitter as a corpus for sentiment analysis and opinion mining // Proc. of the Seventh International Conference on Language Resources and Evaluation (LREC'10). 2010.
9. Al-Kabi M., Al-Ayyoub M., Alsmadi I., Wahsheh H. A prototype for a standard Arabic sentiment analysis corpus // International Arab Journal of Information Technology. 2016. V. 13. N 1A. P. 163–170.
10. Oueslati O., Cambria E., HajHmida M.B., Ounelli H. Review of sentiment analysis research in Arabic language // Future Generation Computer Systems. 2020. V. 112. P. 408–430. https://doi.org/10.1016/j.future.2020.05.034
11. Ghallab A., Mohsen Y., Ali Y. Arabic sentiment analysis: A systematic literature review // Applied Computational Intelligence and Soft Computing. 2020. V. 2020. P. 403128. https://doi.org/10.1155/2020/7403128
12. Duwairi R., Marji R., Sha'ban N., Rushaidat S. Sentiment Analysis in Arabic tweets // Proc. of the 2014 5th International Conference on Information and Communication Systems (ICICS). 2014. P. 1–6. https://doi.org/10.1109/iacs.2014.6841964
13. Bolbol N.K., Maghari A.Y. Sentiment analysis of Arabic tweets using supervised machine learning // Proc. of the 2020 International Conference on Promising Electronic Technologies (ICPET). 2020. P. 89–93. https://doi.org/10.1109/ICPET51420.2020.00025
14. Heikal M., Torki M., El-Makky N. Sentiment analysis of Arabic Tweets using deep learning // Procedia Computer Science. 2018. V. 142. P. 114–122. https://doi.org/10.1016/j.procs.2018.10.466
15. Alhamid M., Alsahli S., Rawashdeh M., Alrashoud M. Detection and visualization of Arabic emotions on social emotion map // Proc. of the International Symposium on Multimedia (ISM). 2017. P. 378–381. https://doi.org/10.1109/ISM.2017.76
16. Al-Thubaity A., Alqahtani Q., Aljandal A. Sentiment lexicon for sentiment analysis of Saudi dialect tweets // Procedia Computer Science. 2018. V. 142. P. 301–307. https://doi.org/10.1016/j.procs.2018.10.494

17. Assiri A., Emam A., Al-Dossari H. Towards enhancement of a lexicon-based approach for Saudi dialect sentiment analysis. *Journal of Information Science*, 2018, vol. 44, no. 2, pp. 184–202. https://doi.org/10.1177/0165551516688143
18. Alqurashi T. Arabic sentiment analysis for twitter data: A systematic literature review. *Engineering, Technology & Applied Science Research*, 2023, vol. 13, no. 2, pp. 10292–10300. https://doi.org/10.48084/etasr.5662
19. Liu B. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. 2nd ed. Cambridge University Press, 2020, 448 p.

17. Assiri A., Emam A., Al-Dossari H. Towards enhancement of a lexicon-based approach for Saudi dialect sentiment analysis // Journal of Information Science. 2018. V. 44. N 2. P. 184–202. https://doi.org/10.1177/0165551516688143
18. Alqurashi T. Arabic sentiment analysis for twitter data: A systematic literature review // Engineering, Technology & Applied Science Research. 2023. V. 13. N 2. P. 10292–10300. https://doi.org/10.48084/etasr.5662
19. Liu B. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions / 2nd ed. Cambridge University Press, 2020. 448 p.

Authors

Ali Benabdallah — PhD, Lecturer, Higher school of Management of Tlemcen, Tlemcen, 13000, Algeria, https://orcid.org/0009-0005-3312-7950, benabdallah.a13@gmail.com

Mohammed Alaeddine Abderrahim — PhD, Lecturer, University of Tlemcen, Tlemcen, 13000, Algeria, sc 57103301000, https://orcid.org/0000-0001-7819-2354, abderrahim.alaa@yahoo.fr

Mohammed Mokri — Researcher, University of Tlemcen, Tlemcen, 13000, Algeria, https://orcid.org/0009-0008-6044-5235, almakri.moh@gmail.com

Авторы

Бенабдаллах Али — PhD, преподаватель, Высшая школа менеджмента Тлемсена, Тлемсен, 13000, Алжир, https://orcid.org/0009-0005-3312-7950, benabdallah.a13@gmail.com

Абдеррахим Мохаммед Алаеддин — PhD, преподаватель, Университет Тлемсена, Тлемсен, 13000, Алжир, sc 57103301000, https://orcid.org/0000-0001-7819-2354, abderrahim.alaa@yahoo.fr

Мокри Мохаммед — исследователь, Университет Тлемсена, Тлемсен, 13000, Алжир, https://orcid.org/0009-0008-6044-5235, almakri.moh@gmail.com

Received 24.07.2024
Approved after reviewing 12.10.2024
Accepted 26.11.2024

Статья поступила в редакцию 24.07.2024
Одобрена после рецензирования 12.10.2024
Принята к печати 26.11.2024