SEBHA UNIVERSITY CONFERENCE PROCEEDINGS, VOL.03 NO. 2 2024
DOI: 10.51984/SUCP.V3I2.3360

وقائع مؤتمرات جامعة سبها
Sebha University Conference Proceedings

Confrence Proceeeding homepage: http://www.sebhau.edu.ly/journal/CAS

# Building Polar-Oriented Libyan Dialect Corpus Using Emoji-Based Lexicon

*Manal M. Athem, Mansour Essgaer, Khamiss M. S. Ahmed, Asma Agaal

Artificial Intelligence Department, Faculty of Information Technology, Sebha University, Sebha, Libya,

**Keywords:**
Polar-oriented lexicon
Emoji symbols
Sentiment analysis
Polar-oriented
Libyan

**ABSTRACT**

The widespread use of social media platforms, such as Twitter, has given rise to research fields focused on analyzing platform content to extract knowledge for decision-making. This study employs natural language processing techniques to construct a Libyan dialect corpus with a focus on polarity, utilizing an emoji lexicon. Initially, tweets in the Libyan dialect were gathered from Twitter and filtered to retain only those containing emoji symbols. Subsequently, exploratory data analysis was conducted to scrutinize the collected tweets, generating a visual statistical interpretation to address various questions. Finally, the polarity of Libyan dialectal tweets was determined through an emoji lexicon-based approach. The results were then assessed by experts, with 80% expressing agreement with the corpus's polarity. The study concludes that emojis play a crucial role in analyzing the sentiment of Libyan youth on Twitter.

# بناء مجموعة اللهجة الليبية ذات التوجه القطبي باستخدام المعجم القائم على الرموز التعبيرية

*منال عبدالعظيم و منصور الصغير و خميس احمد و اسمه اعجال

قسم الذكاء الاصطناعي، كلية تقنية المعلومات ، جامعة سبها، سبها ، ليبيا

الكلمات المفتاحية:
المعجم القطبي المنحى
الرموز التعبيرية
تحليل المشاعر
ذو توجه قطبي
ليبي

الملخص

أدى الاستخدام الواسع النطاق لمنصات التواصل الاجتماعي، مثل تويتر، إلى ظهور مجالات بحثية تركز على تحليل محتوى المنصة لاستخراج المعرفة اللازمة لصنع القرار. تستخدم هذه الدراسة تقنيات معالجة اللغة الطبيعية لبناء مجموعة من اللهجات الليبية مع التركيز على القطبية، وذلك باستخدام معجم الرموز التعبيرية. في البداية، تم جمع التغريدات باللهجة الليبية من تويتر وتصفيتها للاحتفاظ فقط بالتغريدات التي تحتوي على رموز تعبيرية. وبعد ذلك، تم إجراء تحليل البيانات الاستكشافية للتدقيق في التغريدات المجمعة، وإنشاء تفسير إحصائي مرئي لمعالجة الأسئلة المختلفة. وأخيرا، تم تحديد قطبية التغريدات باللهجة الليبية من خلال نهج قائم على المعجم التعبيري. تم بعد ذلك تقييم النتائج من قبل الخبراء، حيث أعرب 80% منهم عن موافقتهم على قطبية الجسم. وخلصت الدراسة إلى أن الرموز التعبيرية تلعب دورا حاسما في تحليل مشاعر الشباب الليبي على تويتر.

## 1. Introduction

Due to the widespread integration of information and communications technology into various aspects of life, social media platforms have gained immense popularity for disseminating news and discussing a wide range of topics. These platforms, such as Twitter [1], have become crucial for facilitating the exchange of opinions and ideas among users.

Twitter, being a micro-blogging platform that enables individuals to express their opinions through text, generates vast amounts of textual data. This data harbors valuable information and knowledge that traditional methods struggle to extract. While conventional methods for textual data analysis exist, they often fall short in exploring the extensive data generated on platforms like Twitter. To address this challenge, many researchers have turned to Natural Language Processing (NLP), a field that has achieved significant success in various domains, including education, healthcare, and agriculture [2]. NLP refers to the ability of a computer program to comprehend human language as it is spoken and written.

Sentiment Analysis (SA), a branch of Natural Language Processing, focuses on textual data and involves determining the negative and positive polarity of users' opinions and feelings [3]. Sentiments can vary widely among individuals in terms of negative and positive intensity, posing a challenge in accurately determining feelings. Emojis, which often accompany text, add another layer of complexity to this challenge [1]. Emojis are commonly used to convey hidden and ambiguous feelings, making them increasingly valuable in SA. A previous study [1] analyzed 13 European languages using a

* Manal M. Athem:
E-mail addresses: mana.athem@fit.sebhau.edu.ly, (M. Essgaer) man.essgaer@sebhau.edu.ly, (K. M. S. Ahmed) km.ahmed@sebhau.edu.ly, (A. Agaal) asma.agaal@sebhau.edu.ly

*Article History : Received 19 June 2024 - Received in revised form 16 September 2024 - Accepted 06 October 2024*

lexicon-based approach, resulting in a comprehensive lexicon containing sentiments related to emojis, applicable in NLP and linguistic research. The use of emojis enhances the continuity of individual communication, improves relationship quality, and strengthens emotional communication between users [4].

Various approaches have been employed to define polarity, including dictionaries or lexicon-based approaches, descriptive or predictive machine learning approaches, and hybrid approaches. SA relying on emoji sentiments is a recent research focus. However, text-based analysis methods may fall short in sentiment analysis if the emojis within the text are not considered. To the best of the author's knowledge, few studies related to Libya discuss SA using Libyan dialectal sentences [5]. The mentioned study examined tweet text without taking into account the emojis present in the tweets. Therefore, this study collects Libyan dialectal tweets from Twitter that contain emoji symbols, conducts exploratory analysis on the collected tweets, and ultimately constructs a Libyan dialectal emoji-based corpus. This corpus is considered the first to address emoji-related issues using the lexicon-based approach within the context of the Libyan dialectal language.

The remaining sections of this paper are structured as follows: Section 2 provides an overview of related work, Section 3 details the materials and techniques employed in the study, Section 4 covers the experiments and subsequent discussion, and finally, Section 5 concludes the paper.

## 2. Literature Reviews

Given the distinctive grammatical and structural characteristics of Arabic, Natural Language Processing (NLP) encounters significant challenges when applied to this language. Additionally, the scarcity of references, studies, and corpora, particularly those addressing dialectical language, further complicates the NLP landscape. Achieving robust results in NLP applications often requires a substantial raw corpus, and existing discussions typically revolve around Modern Standard Arabic, officially employed in media like TV and newspapers. Despite Arabic being spoken by approximately two hundred million people, numerous local dialects, including those in Libyan regions, present further complexities [6].

Considering the dearth of studies specifically focusing on the Libyan dialect, alternative research endeavors are explored in this section. In a study conducted by [7], a lexical-based approach was employed to ascertain polarity. The collected text was initially translated into a reference language, and a polar-oriented dictionary assigned degrees of sentiment to each text. Another study by [8] also adopted the lexicon-based approach for determining sentiment polarity, enhancing it by incorporating negation statements.

From a different perspective, the increasing use of emojis on various social media platforms, such as Twitter, has garnered attention. Emojis offer the advantage of directly expressing emotions irrespective of the written language. In the study conducted by [9], sentences containing emojis were found to convey sentiments more explicitly and comprehensibly compared to sentences without emojis. Expanding on this idea, [10] proposed extending binary sentiment classification approaches to include multi-way emotions classification, demonstrating that emojis significantly enhance precision in recognizing diverse emotions. In a lexicon-based approach to analyzing Arabic sentiments, [6] highlighted the challenge of requiring extensive data for models to achieve higher accuracy.

Furthermore, a study by [11] delved into the gender-specific use of emojis in textual communication. Employing a questionnaire to gather data from 30 participants, the researchers assumed that women used emojis more frequently than men, only to discover that the results contradicted this assumption. These diverse studies collectively underscore the multifaceted nature of sentiment analysis and the evolving role of emojis in conveying emotions within the Arabic language context.

## 3. Material and method

The proposed study uses a framework derived from the general Data Mining (DM) methodology as shown in Fig.1, represented in the following steps: data collection stage, exploratory and processing stage, and polarity identifications stage using the lexicon-based approach, and lastly, the evaluation of the results.

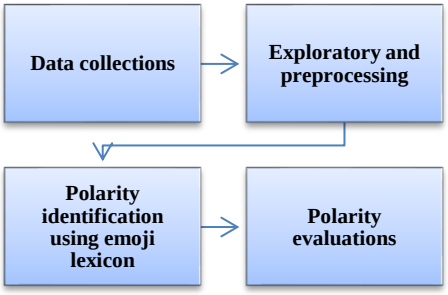


**Fig. 1.** The Proposed Framework

Data serves as a fundamental resource for social networking platforms, and the data access policies of each site vary for the research community. Twitter, with relatively fewer data access restrictions, stands out as one of the optimal platforms for research purposes. Consequently, Twitter was utilized to gather tweets related to Libya, employing predefined Libyan dialect keywords outlined in Table (1). The data collection stage comprises three phases:

1. **Collecting the Libyan Dialectal Tweets Corpus**: This involved the acquisition of tweets based on the specified keywords from January to March 2021.
2. **Filtering the Corpus**: To eliminate non-Libyan-related tweets and mitigate the presence of other Arabic dialects resulting from the use of similar keywords, a filtering process was employed. Content filtering involves refining data to keep only relevant information. It typically includes:
   - ***Keyword filtering***: Keeping data with specific, relevant words.
   - ***Language/Dialect*** detection: Identifying and removing irrelevant dialects or languages.
   - ***Topic filtering***: Ensuring the data aligns with the subject of interest.
3. **Selecting Tweets with Emoji Symbols**: The final phase involved the selection of tweets containing emoji symbols, as these symbols play a crucial role in sentiment analysis.

**TABLE I. PREDEFINED LIBYAN DIALECT KEYWORDS TO COLLECT THE CORPUS**

| | | | | | | |
|---|---|---|---|---|---|---|
| ما جاش | العشيه | انخدشت | تركينه | الدخلاني | رقدت | الفترينه |
| حطيت | يشبح | زوز | حوش | حوايج | الكوفينو | بنزينه |
| متاعي | باهي | هنايا | اوني | هادو | غدوه | قعمز |
| اوخيك | زعمه | اماله | خبط | القوانتي | تزداني | عزومة |
| فياق | بكرج | روشن | كبوطي | ضي | قداش | مليح |
| ماظنش | هلية | مكسد | ماكلة | صونية | مصقع | نتلاقو |
| طاسه | حقاني | | | | | |

Table (2) displays a sample of five rows from the filtered corpus, which comprises a total of thirty thousand rows, providing the dataset for the subsequent experiments

**TABLE II. DEMONSTRATES PART OF THE COLLECTED CORPUS**

| Tweet date | Tweet text |
|---|---|
| Sat Jan 30 15:31:36 2021 | @USERS لو نقولك اني مره رقدت يومين متو... |
| Sun Feb 21 18:42:56 2021 | الي واجعتني رقدت من التعب و ما حضرتهاش لاكن بر... |
| Mon Feb 01 20:49:15 2021 | اسمعوا لو حطيت هذا بس 😭 😭 😭 😭 يعني اسلك |
| Fri Jan 15 13:29:27 2021 | @USERS روشن جانم 😍 😁 |
| Thu Jan 14 01:13:30 2021 | @USERS قاعده فالفترة متاع كلمة " كنك " حا تبكيني ... |

The second phase involves the exploratory and pre-processing stage, which entails a comprehensive understanding of the collected corpus and its contents. This stage involves exploring all features of the corpus to identify any instances of noise, missing values, or outliers. Utilizing various statistical methods, charts, and graph representations, this step aims to test hypotheses, verify assumptions, and enhance the overall data quality. It is considered

a crucial step in any Data Mining (DM) methodology, as the quality of the data significantly influences the accuracy of polarity determination.

Following the exploratory and pre-processing stage, the subsequent phase focuses on identifying polarity. This is achieved by assigning sentiment scores to each emoji in consideration using an emoji lexicon. The polarity of emojis within each tweet is then aggregated based on their positive and negative scores. The lexicon utilized in this study, as depicted in Table (3), was proposed by [1]. The table includes the emoji symbol, positive and negative scores, neutral score, and the classification of each emoji symbol. This lexicon serves as the basis for determining the sentiment polarity associated with emojis in the collected tweets.

**TABLE III. THE LEXICON USED TO ADDRESS THE TWEETS EMOJI**

| Image | Negative | Neutral | Positive | Description of emoji | Classification |
|---|---|---|---|---|---|
| 😂 | 0.285 | 0.468 | 0.221 | Face With Tears of Joy | Emoticons |
| ❤ | 0.166 | 0.790 | 0.746 | Heavy Black Heart | Dingbats |
| ♥ | 0.272 | 0.693 | 0.657 | Black Heart | Miscellaneous |
| | | | | Suit | Symbols |
| 😍 | 0.219 | 0.729 | 0.678 | Smiling Face with Heart-Shaped Eyes | Emoticons |
| 😭 | 0.220 | 0.343 | -0.093 | Loudly Crying Face | Emoticons |

Finally, evaluating the polarized tweets is based on an evaluation form distributed among researchers in the same field; this step ensures the results' quality.

## 4. Results and discussions

The results of this study are presented in two distinct stages: the initial stage involves exploratory data analysis, wherein the corpus is scrutinized to extract valuable information and knowledge. The subsequent stage focuses on determining the polarity of the collected tweets using the emoji lexicon.

In the first stage, a series of questions were posed, leveraging both qualitative and quantitative features of the collected corpus. Qualitative features include nominal and numerical aspects, such as the tweet's timestamp, the source of the tweeting device, the user's profile name, favorite counts, and retweet counts.

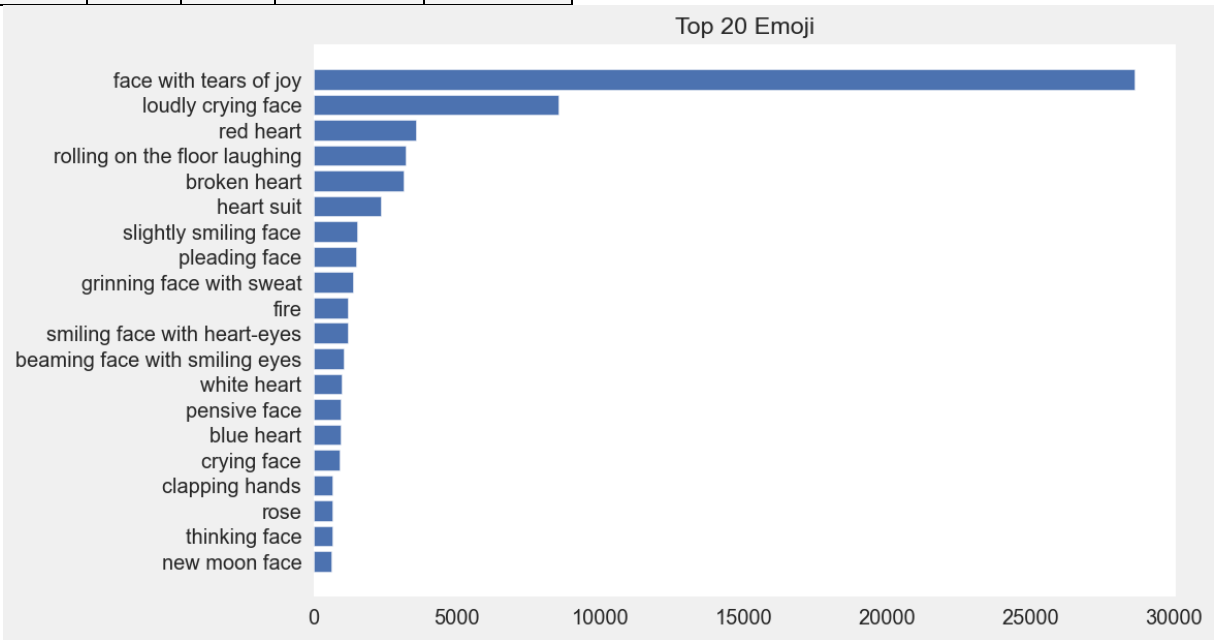


**Fig. 2.** Shows the frequency of emojis in tweets

In the subsequent stage, the polarity of the tweets is determined using the emoji lexicon. The lexicon-based Sentiment Analysis (SA) has demonstrated its effectiveness in classifying tweets based on emojis in terms of their polarity. However, it is worth noting that the lexicon SA faced challenges with certain non-modern emoji faces, as illustrated in Table (4).

Examining Table (4), it is evident that some tweets received positive scores, primarily due to the positive sentiments associated with the majority of emojis in those tweets. Conversely, some tweets received negative scores owing to the presence of negatively associated emojis. The sentiment score in certain cases is influenced by a single emoji, as

One of the key questions aimed to reveal fundamental insights, specifically identifying the most frequently used emojis in tweets, as illustrated in Fig.2. The figure 2 highlights that the four most prevalent emojis were positively associated, including the face with tears of joy, loudly crying face, red heart, and rolling on the floor laughing. This observation suggests that in 2021, social media posts made by Libyans, particularly the youth, predominantly conveyed positive sentiments.

exemplified by tweet number 2. In contrast, in other instances, the sentiment score is determined by the aggregation of scores from several emojis, as observed in tweets number 1, 3, 4, and 5. This nuanced approach reflects the varying impact and influence of emojis on the overall sentiment expressed in the tweets.

**TABLE IV. DESCRIBES THE DATABASE OF EMOJIS**

| | Tweet texts | Emojis in tweet | Sentiments score |
|---|---|---|---|
| 1 | ادعمو اختكم عبير العالميه ان زعل واحد يرضونك ... | 💛 💙 💐 🤚 | 72% |
| 2 | اه لما تجيني اشعارات من شخصي يفز قلبي😳 | 😳 | -52% |
| 3 | بطلت بعقل خلاص من المداخلات صارت تجيني حسابات... | 😂 💔 🙄 | 5% |
| 4 | معا من يتحاربو هادو 🤣🤣🤣🤣🤣🤣🤣 | 🤣🤣🤣🤣🤣 | 26% |
| 5 | فعالية وهدية 💸 💰 بعد شوية خليك قريب🟥 | 💰 💸 🟥 | 20% |

Additionally, the sentiment scores fall within the range of 1 to -1, where a score of 1 indicates a perfectly positive sentiment and a score of -1 reflects a perfectly negative sentiment. Notably, the study

encountered a challenge in determining the polarity of certain tweets due to the presence of emojis not recognized in the lexicon. Consequently, these tweets were excluded from the corpus to ensure the accuracy of the analysis.

In the evaluation of the obtained results, an assessment form was designed and distributed among researchers with expertise in sentiment analysis. The results revealed a high level of agreement between the lexicon-generated scores and the assessments of the researchers. As depicted in Fig.3, the agreement rate approached almost 80%, underscoring the reliability and effectiveness of the lexicon-based approach in determining the sentiment of Libyan dialectal tweets

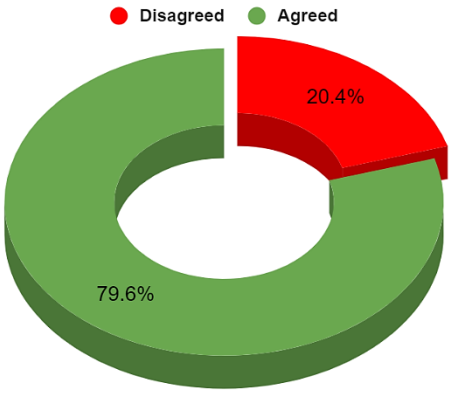


**Fig. 3.** Lexicon result evaluations based on the opinion of researchers in the sentiment analysis field

Based on the findings, the study draws the conclusion that analyzing textual sentiment is a complex task, and managing data proves to be a resource-intensive process demanding significant effort and time, particularly when dealing with large datasets. The study's primary emphasis was on collecting data specific to the Libyan dialect, followed by a comprehensive exploration and analysis, culminating in the determination of polarity within the collected texts using the Emoji Dictionary. This underscores the intricate nature of sentiment analysis, especially in linguistic contexts like the Libyan dialect, and highlights the importance of thorough data curation and analysis methodologies.

## 5. Conclusion and Recommendations

This study underscores the increasing significance of micro-blogging platforms like Twitter for sentiment analysis, as evidenced by its outcomes. The research reveals a critical gap in sentiment analysis studies pertaining to Arabic sources, particularly within the scope of the Libyan dialect. Consequently, the creation of a dedicated Libyan dialect corpus is identified as imperative, offering benefits for dialect identification, translation, and related tasks. Emojis extracted from the corpus prove to be essential indicators in determining text polarity, emphasizing their crucial role in conveying sentiments. The study stands out for contributing a valuable asset to the research domain – the second Libyan dialect corpus and the inaugural application of the emoji lexicon for text polarity determination. Looking ahead, the study advocates for future research endeavors to consider geographical location data for more nuanced analyses, develop predictive models for forecasting sentiment based on emojis in the Libyan dialect, and explore lexicons that accommodate modern emojis to enhance sentiment identification accuracy. These recommendations seek to advance the landscape of sentiment analysis research within the distinctive context of the Libyan dialect.

## References

[1]- Boutet, I., et al., *Emojis influence emotional communication, social attributions, and information processing*. 2021. **119**: p. 106722.

[2]- Doliashvili, M., M.-B.C. Ogawa, and M.E. Crosby. *Understanding Challenges Presented Using Emojis as a Form of Augmented Communication*. in *Augmented Cognition. Theoretical and Technological Approaches: 14th International Conference, AC 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part I 22*. 2020. Springer.

[3]- Omar, A., M. Essgaer, and K.M. Ahmed. *Using machine learning model to predict libyan telecom company customer satisfaction*. in *2022 International Conference on Engineering & MIS (ICEMIS)*. 2022. IEEE.

[4]- Bai, Q., et al., *A systematic review of emoji: Current research and future perspectives.* 2019. **10**: p. 2221.

[5]- Alhammi, H.A., K.J.I.J.o.C.T. Haddar, and Engineering, *Building a libyan dialect lexicon-based sentiment analysis system using semantic orientation of adjective-adverb combinations.* 2020. **12**(6): p. 145-150.

[6]- Sherif, S.M., et al., *Lexicon annotation in sentiment analysis for dialectal Arabic: Systematic review of current trends and future directions.* 2023. **60**(5): p. 103449.

[7]- Chakraborty, K., S. Bhattacharyya, and R.J.I.T.o.C.S.S. Bag, *A survey of sentiment analysis from social media data.* 2020. **7**(2): p. 450-464.

[8]- Albesher, A.S., O.B.J.I.J.o.M. Rabie, Semantics, and Ontologies, *A survey study on Arabic WordNet: baring opportunities and future research directions.* 2020. **14**(4): p. 290-305.

[9]- Jones, L.L., et al., *Sex differences in emoji use, familiarity, and valence.* 2020. **108**: p. 106305.

[10]- Wolny, W., *Emotion analysis of twitter data that use emoticons and emoji ideograms.* 2016.

[11]- Persson, N., *Analysis of Emoji Usage: Differences in Preference and Function Across Genders*. 2019.