Lebanese Arabizi Tweets Sentiment Analysis

M. Benabid*
* Department of Computer Science and Mathematics, Lebanese American university, Beirut, Lebanon
manelroumaissa.benabid@lau.edu

***Abstract*—Sentiment analysis is a widely used technique to detect the opinion of the population towards a certain event, product, or personality. Its use has spread widely in the last couple of years due to its benefits in marketing, economics, politics, and many more. Arabizi is a form of writing Arabic in Roman letters. There are many ways of writing Arabizi because it depends on the dialect of the person writing. Lebanese Arabic is a quite unique form of writing Arabizi due to it being a mix of Arabic, English, French, and Armenian. Other forms of Arabizi, such as Algerian Arabizi, are also unique to it being a mix of Arabic, Amazigh (or Berber), French, English, Spanish, and Italian. In our study, we propose different models for classifying the sentiment of Lebanese Arabizi text through tweets. We relied on an already existing paper. Our work outperformed the already existing work with an accuracy of 75% and an F1 score of 0.75.**

***Index Terms*—Arabizi, Tweets, Sentiment Analysis, Machine Learning, Natural Language Processing, Deep Learning.**

I. Introduction

Sentiment analysis, also known as opinion mining or emotion AI, is a field within natural language processing (NLP) that uses various techniques to analyze text and identify the underlying sentiment, opinion, or emotion. It aims to understand the subjective aspects of language, beyond just factual information.

The process uses a combination of: **Computational linguistics**: Analyzing the structure and meaning of language. **Text analysis**: Identifying keywords, phrases, and sentence structure. **Machine learning**: Training algorithms on large datasets of text and sentiment labels.

Through these techniques, sentiment analysis can classify the overall sentiment of a piece of text as positive, negative, or neutral. It can also delve deeper to identify specific emotions, such as anger, joy, or sadness.

Sentiment analysis has a wide range of applications across various industries. Here are some key reasons why it's used:

1. **Brand Monitoring**: Businesses can use sentiment analysis to monitor online conversations about their brand, products, or services. This allows them to understand public perception, identify potential issues, and improve customer experience.

2. **Market Research**: Businesses can analyze social media posts, customer reviews, and other online data to understand customer sentiment toward their products, competitors, and industry trends. This information can guide them in making informed decisions about marketing, product development, and pricing.

3. **Social Media Analysis**: Sentiment analysis can help analyze the overall sentiment of a social media campaign, track the effectiveness of marketing messages, and identify influencers who are promoting positive or negative sentiment.

4. **Customer Service**: Companies can analyze customer feedback and complaints to identify areas for improvement and provide better customer service. Sentiment analysis can also help identify upset customers who need immediate attention.

5. **Political Analysis**: Sentiment analysis can be used to analyze the public's opinion on political candidates, policies, and events. This information can be valuable for politicians and political analysts to understand public sentiment and adjust their strategies accordingly.

6. **Healthcare**: Sentiment analysis can be used to analyze patient feedback and identify areas where healthcare providers can improve their services. It can also be used to identify patients who are at risk of depression or other mental health issues.

7. **Finance**: Sentiment analysis can be used to analyze financial news and reports to predict market trends. It can also be used to identify potentially fraudulent activity.

Arabizi, also known as Arabish, Arabglizi, or the Arabic chat alphabet, is a fascinating linguistic phenomenon born in the digital age. It's an informal writing system used primarily by young Arabs for online communication, particularly in texting and social media.[1]

Arabizi is essentially a transliteration of Arabic into Latin letters and Arabic numerals. This allows for faster typing on keyboards designed for Latin alphabets, which are often the default on mobile devices.[2] [3] Specific symbols are used to represent sounds unique to Arabic that don't exist in English. For example:

ح (Haa): Represented by the number "7" due to its similar shape.

ع ('ayn): Represented by the number "3" due to its resemblance when flipped.

ء (hamza): Represented by an apostrophe (').

Several factors contribute to the popularity of Arabizi:

Convenience: Typing in Latin characters is much faster than using the Arabic script on mobile devices. Digitalization: Arabizi reflects the increasing digital communication among young Arabs, who are largely fluent in both Arabic and English. Informal Communication: Arabizi is considered a casual and informal way of writing, perfect for online chats and social media interactions. Trendy and Evolving: Arabizi constantly evolves, incorporating new slang terms, emojis, and abbreviations, making it a dynamic and exciting language.

The Arabizi version depends on the local dialect. For example, the Algerian Darija is different from the Syrian and Lebanese Arabic. Lebanese Arabizi is a mix of Arabic, French, Armenian, and English [4] which makes it a complex language to understand and categorize.

Large Language Models and their recent usage in NLP:

Large Language Models (LLMs) are a type of artificial intelligence (AI) that has revolutionized the field of Natural Language Processing (NLP) in recent years. LLMs are trained on massive amounts of text data, allowing them to learn complex patterns and relationships in human language. This enables them to perform a wide range of NLP tasks with impressive accuracy, including Sentiment analysis among other tasks such as text generation, translation,question answering, and many more.

The work of [5] carried out an investigation to assess the utilization of Arabizi on Twitter and to create automated tools for identifying Arabizi in multilingual data streams. Twitter data from Lebanon and Egypt was gathered, and the percentage of each language, especially Arabizi, was reported, offering valuable insights for researchers analyzing natural text in the Arab region. To accomplish this, they developed an Arabizi identification classifier by annotating sample datasets and extracting features using Langdetect, a language detection library. The study achieved an average classification accuracy of 93% and 96% for Lebanon and Egypt datasets respectively. The classifier for Arabizi identification depended on features at the sentence level, and the authors recommended enhancing it by extracting features at the word level from the text.

The research also sought to progress Arabic Natural Language Processing (NLP) studies by enabling the examination of social media data without the necessity to filter out intricate or less common languages. The authors emphasized the significance of identifying Arabizi in multilingual social media data, as it brings researchers a step closer to tackling sentiment analysis for Arabic, encompassing both Non-Standard Arabic (NSA) and Arabizi.

II. Related Work

The authors of [6] performed sentiment analysis on Twitter data through distant learning to obtain sentiment information, constructed models utilizing Naive Bayes, MaxEnt, and Support Vector Machines (SVM), and experimented with a Unigram, Bigram model in combination with parts-of-speech (POS) features. Introduced POS-specific prior polarity features and a tree kernel to enhance sentiment analysis on Twitter. The outcomes revealed that these novel features, in addition to previously suggested features, achieved comparable performance to the state-of-the-art baseline, surpassing its effectiveness.The study of [7] aimed at classifying tweets into positive, negative, and neutral categories led to the creation of a suggested system for analyzing sentiment in Twitter data. The process included retrieving tweets, preprocessing the extracted data, employing parallel processing, incorporating a sentiment scoring module, and generating output sentiments. The research also introduced an innovative hybrid method that combines corpus-based and dictionary-based techniques to ascertain the semantic orientation of sentiment words in tweets. Furthermore, the study emphasized the utilization of machine learning algorithms for predicting sentiment on Twitter, resulting in superior performance compared to three models using Weka.

The work of [8] investigated the application of supervised learning to attribute sentiment labels to Arabizi tweets. The research demonstrated the efficacy of Support Vector Machines (SVM) in the sentiment classification of Arabizi tweets and highlighted the influence of different preprocessing techniques on classification accuracy. The study outcomes indicated that the SVM classifier exhibited superior performance compared to the Naive Bayes classifier in the sentiment classification of Arabizi tweets. Additionally, it was observed that eliminating stopwords and substituting emoticons with their corresponding words did not significantly enhance the classification accuracies for Arabizi data. The authors of [9] developed a novel Arabizi language model named BAERT and assessed its performance in sentiment analysis tasks using LAD and SALAD datasets. The findings indicated that BAERT exhibited better performance than the multilingual BERT model specifically in sentiment analysis tasks resembling Tunizi and Egtptizi sentiment analysis. The study also observed that initiating the training of a Transformer model from scratch is computationally more demanding compared to commencing with pre-trained weights. The authors of [10] proposed a method for analyzing Arabizi, involving the automatic categorization of sentiments in Arabizi messages into positive and negative categories. The approach comprises the following key steps:

1) Automatic construction of an Arabic sentiment lexicon.
2) Automatic annotation of Arabic messages.
3) Arabizi transliteration.
4) Sentiment classification of Arabic messages.

The researchers employed shallow machine learning algorithms such as Support Vectors Machine (SVM) and Naive Bayes (NB) for corpus validation. Results demonstrated that the Naive Bayes algorithm outperformed other algorithms, achieving the highest F1-score of up to 78% for manually transliterated datasets and 76% for automatically transliterated datasets. The authors also mentioned ongoing efforts aimed at enhancing the transliterator module and annotated sentiment dataset.

The authors of [11] introduced a supervised method for sentiment analysis in the Algerian dialect written in the Latin script using Arabizi. The researchers gathered and annotated three datasets through crowdsourcing, evaluating the impact of classifiers, presentation types, and innovative contributions in the preprocessing phase, including the removal of vowels. The approach yielded an F1-score of 87% and an accuracy of 83%, with SVM demonstrating superior performance among classifiers. Through preprocessing, they managed to enhance the F1-score of SVM by 9.20%, signifying a noteworthy improvement and reinforcing the significance of their initial assumptions. In summary, their results exhibit promising performance and underscore the positive influence of the proposed preprocessing techniques.

The purpose of the paper of [5] was to investigate the utilization of Arabizi on Twitter and develop tools for automatically detecting Arabizi in diverse streams of data. The authors

gathered Twitter data from Lebanon and Egypt with the aim of offering valuable insights for researchers analyzing natural text in the Arab region. Specific goals encompassed training an Arabizi identification classifier, extracting features using established language detection libraries, and contributing to the advancement of Arabic Natural Language Processing (NLP) research by enabling analysis of social media data without the necessity to filter out complex or minority languages. The results of the study included the following key findings:

1) The Arabizi identification classifier yielded an average classification accuracy of 93% and 96% for the datasets from Lebanon and Egypt, respectively.
2) The research revealed that the percentage of Arabizi usage in Twitter data in both Lebanon and Egypt was comparatively lower than reported by other researchers in mobile messaging. Nonetheless, a notable portion of a country's Twitter data—4% or 5.7%—was identified as Arabizi, suggesting a significant volume of data potentially containing valuable information.
3) Unidentified Arabizi tweets were predominantly observed in tweets from Lebanon, written in both English and Arabizi. The study emphasized the importance of developing Natural Language Processing (NLP) resources to effectively identify, analyze, and process Arabizi data.

The primary objective of [12] was to present a dedicated sentiment analysis approach tailored for Arabic and its dialects, with a specific emphasis on the Algerian dialect. The study aimed to address both Arabic script and Arabizi, utilized for the Algerian dialect while comparing various word embedding models and deep learning classifiers for sentiment analysis. Additionally, the paper sought to tackle the challenges linked to sentiment analysis in Arabic and its dialects, and it aimed to suggest future directions for enhancing the proposed approach. The study yielded highly promising results, achieving F1 scores of up to 89% for extrinsic experiments utilizing CNN. This performance surpassed existing research literature by as much as 25%. The research underscored the effectiveness of transliteration in Arabizi sentiment analysis and put forth recommendations for future enhancements. These include the development of a transliteration system based on a corpus-based approach, the enrichment of parallel corpora, an extension of the constructed lexicon using Word2vec, and the proposition of classifiers that amalgamate different models. Additionally, the study emphasized the significance of reviewing automatically constructed resources and suggested novel techniques and approaches for constructing resources with minimal effort.

III. Methodology

In this section, we detail the methodology employed in our study, outlining the steps taken to collect, preprocess, and analyze the data. The methodology encompasses the utilization of both traditional machine learning classifiers, such as Naive Bayes and Logistic Regression, as well as deep learning classifiers and a BERT-based model.

A. Data Collection

The dataset utilized in this study consists of Lebanese Arabizi Tweets that were scraped using the Twitter API. These Tweets were published between January 2017 and April 2020 and they all contained keywords such as "chou, kamen, chwei,.." which are common words in the Lebanese Dialect and often used in everyday conversations. The tweets were also filtered based on the location and cleaned to remove all spam, images, duplicates, and noise. Data was collected and annotated by [4] and is publicly accessible at Kaggle. The final dataset includes tweets, sentiment, and highlight (emotion).

B. Data Preprocessing

Prior to analysis, the raw data underwent a series of preprocessing steps. These included converting the tweets to lowercase, removing punctuation, mentions, hashtags, URLs, standalone numbers, measurements, and timings, and removing stop words. Since the Lebanese Arabizi is a mix of Arabic-Lebanese, English, and French, both English and French stopwords were removed in addition to a manually curated list of Lebanese Arabic stopwords (e.g. include "chou, aw, w, hay, lech, etc."). Moreover, exaggerations were reduced (e.g. "whyyyy" was reduced to "why") and sound effects were replaced by their equivalent meaning. For example, "hahahahha" was replaced by "laughter". Finally, some emojis were replaced by their meaning if they were significant to the sentiment prediction. For example, happy-face emojis, winking emojis, and so on were replaced by the phrase "happy face smiley" and sad, crying emojis were replaced by the phrase "frown angry sad pouting", other non-important emojis such as "eyes" and those representing animals were discarded of as they do not reflect the sentiment of the tweet. The purpose of these steps was to remove noise from the data that can interfere with the classification process. Stopwords are not correlated with a certain sentiment so it is recommended to remove them so they don't interfere with the classification process. Hashtags, URLs, and mentions do not reflect a person's sentiment so they can be removed.

C. Machine Learning Classification Approaches

Seven machine learning models were implemented: Multinomial Naive Bayes, Logistic Regression, Gradient Boosting, Random Forest, Decision Trees, Support Vector Machines (linear, polynomial (degree =2), RBF, sigmoid), and a Recurrent Neural Network (RNN) model using bidirectional LSTMs.

We implemented the base models and models with hyperparameter tuning using RandomizedSearchCV/GridSearchCV. The RNN parameters were the same as used by [4].

D. BERT Model

The tokenizer used was BertTokenizer from the transformers library. We used the 'bert-base-uncased' pre-trained model, we also used the AdamW optimizer with a learning rate of 1e-5. We trained the model for 10 epochs and saved each model. The final model (epoch 10) was used for testing on the test set.

IV. Experimental Results

The data was split into 80%-20% for training and testing respectively.

Table I shows the test results of the base (before hyperparameter tuning) machine learning models

Algorithm	Accuracy	Precision	Recall	F1
Multinomial Naive Bayes	0.70	0.70	0.60	0.65
Logistic Regression	0.72	0.67	0.74	0.70
Gradient Boost	**0.75**	0.71	**0.78**	**0.74**
Random Forest	0.72	**0.72**	0.66	0.69
Decision Trees	0.61	0.58	0.60	0.59
SVM polynomial	0.68	0.76	0.44	0.56
SVM linear	0.70	0.66	0.71	0.69
SVM sigmoid	0.50	0.45	0.51	0.48
SVM RBF	0.69	0.66	0.69	0.67

TABLE I
Base ML Models test results

Table II shows the best parameters obtained for the logistic regression hyperparameter tuning.

Parameter	Value
C	0.6158
max_iter	100
penalty	l1
solver	liblinear

TABLE II
Logistic Regression Best Parameter

Table III shows the best parameters of the random forest model after hyperparameter tuning.

Parameter	Value
n_estimators	1200
min_samples_split	5
min_samples_leaf	2
max_features	sqrt
max_depth	20
bootstrap	True

TABLE III
Random Forest Best Parameters

Table IV shows the parameters of the decision tree model after hyperparameter tuning.

Parameter	Value
criterion	gini
min_samples_split	8
min_samples_leaf	3
max_depth	5

TABLE IV
Decision Trees Best Parameters

Table V shows the parameters of the SVM model after hyperparameter tuning.

Table VI shows the test parameters of the machine learning models after hyperparameter Tuning

Fig 1 shows the training accuracy and validation accuracy (left) and the training and validation loss (right).

Parameter	Value
C	1
gamma	scale
kernel	linear

TABLE V
SVM Forest Best Parameters

Algorithm	Accuracy	Precision	Recall	F1
Logistic Regression	0.72	0.67	**0.74**	0.70
Random Forest	**0.75**	**0.74**	0.72	**0.73**
Decision Trees	0.66	0.68	0.49	0.57
Best SVM	0.70	0.66	0.71	0.69

TABLE VI
Hyperparameter Tuned ML Models test results

The BERT model achieved a test accuracy of 74.58% and an F1 score of 0.75.

V. Conclusion

We note that the best model for predicting the sentiment was Gradient Boost with an accuracy of 0.75, precision of 0.71, recall of 0.78, and F1 score of 0.71. We highlight that the results of the gradient boost were on par with the tuned random forest and BERT model. Overall, we outperformed the results of [4] by far.

The deep learning model performed poorly compared to the other machine learning model. It is assumed that it's because we do not have enough data for the model to learn from. Hence, it reached a test accuracy of 0.5, we note that the model was unable to correctly classify any observation of class 1 (Positive).

The BERT model performed quite well it was able to correctly classify 84 out of 119 negative cases and 92 out of 117 positive cases with an overall accuracy of 74.58% and 0.75 F1 score.

VI. Discussion

In this section, we will discuss the obtained results.

First, we highlight the fact that we used the BERT-uncased model that was trained on English corpus, we did that because there is no pre-trained LLM model on Lebanese Arabizi so that poses a limitation on our classification task. A good work that can be done in the future is to train a brand new model on Lebanese Arabizi text alone, which will certainly provide

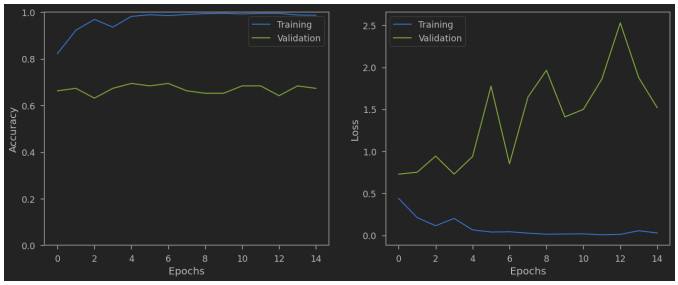

Fig. 1. Deep Learning model training results

way better results than the BERT-base-uncased that was not trained to detect the Lebanese dialect. Moreover, we notice that the deep learning model performed poorly compared to other machine learning models, which makes us think that the reason for that is the lack of data. It is worth noting that deep learning models require a huge amount of data to be able to generalize to unseen data, otherwise, the model will over-fit the training data, and that is the case at hand. The highest achieved accuracy was 50% which is very low compared to other Machine Learning models. We also highlight that the model only ran for 14 epochs, due to not having enough data for the model to train for longer. As for the machine learning model's performance, the best models were the Gradient Boost and the fine-tuned Radom Forest, they outperformed all other models (including LSTM and BERT) with an accuracy of 0.75 and F1 score of 0.75. But it is worth noting that the ML models' performance improved significantly after the fine-tuning step. For example, the decision tree model increased from 0.61 to 0.66. On another note, the GridSearch parameter tuning on the SVM model found that the best model to predict the sentiment of Arabizi text is the linear model with a good accuracy of 0.72 and an F1 score of 0.69.

REFERENCES

1 Yaghan, M. A., "" arabizi": A contemporary style of arabic slang," *Design Issues*, vol. 24, no. 2, pp. 39–52, 2008.

2 Darwish, K., "Arabizi detection and conversion to arabic," *arXiv preprint arXiv:1306.6755*, 2013.

3 Allehaiby, W. H., "Arabizi: An analysis of the romanization of the arabic script from a sociolinguistic perspective." *Arab World English Journal*, vol. 4, no. 3, 2013.

4 Raïdy, M. and Harmanani, H., "A deep learning approach for sentiment and emotional analysis of lebanese arabizi twitter data," in *International Conference on Information Technology-New Generations*. Springer, 2023, pp. 27–35.

5 Tobaili, T., "Arabizi identification in twitter data," in *Proceedings of the acl 2016 student research workshop*, 2016, pp. 51–57.

6 Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R. J., "Sentiment analysis of twitter data," in *Proceedings of the workshop on language in social media (LSM 2011)*, 2011, pp. 30–38.

7 Sahayak, V., Shete, V., and Pathan, A., "Sentiment analysis on twitter data," *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*, vol. 2, no. 1, pp. 178–183, 2015.

8 Duwairi, R. M., Alfaqeh, M., Wardat, M., and Alrabadi, A., "Sentiment analysis for arabizi text," in *2016 7th International Conference on Information and Communication Systems (ICICS)*. IEEE, 2016, pp. 127–132.

9 Baert, G., Gahbiche, S., Gadek, G., and Pauchet, A., "Arabizi language models for sentiment analysis," in *Proceedings of the 28th international conference on computational linguistics*, 2020, pp. 592–603.

10 Guellil, I., Adeel, A., Azouaou, F., Benali, F., Hachani, A. E., and Hussain, A., "Arabizi sentiment analysis based on transliteration and automatic corpus annotation," in *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis*, 2018, pp. 335–341.

11 Chader, A., Lanasri, D., Hamdad, L., Belkheir, M. C. E., and Hennoune, W., "Sentiment analysis for arabizi: Application to algerian dialect." in *KDIR*, 2019, pp. 475–482.

12 Guellil, I., Adeel, A., Azouaou, F., Benali, F., Hachani, A.-E., Dashtipour, K., Gogate, M., Ieracitano, C., Kashani, R., and Hussain, A., "A semi-supervised approach for sentiment analysis of arab (ic+ izi) messages: Application to the algerian dialect," *SN Computer Science*, vol. 2, pp. 1–18, 2021.