PoliTun: Tunisian Political Dataset for Detecting Public Opinions and Categories Orientation

Chayma Fourati
National School of Comupter Science / Tunisia
chaymafourati12@gmail.com

Roua Hammami
Higher Institute of Multimedia Arts / Tunisia
roua.hammami@yahoo.com

Chiraz Latiri
Faculty of Sciences of Tunis, Tunis EL Manar University / Tunisia
chiraz.latiri3@gmail.com

Hatem Haddad
Manouba University / Tunisia
haddad.hatem@gmail.com

Abstract

In recent years, social media platforms and online forums have become essential sources of political discourse, reflecting the ever-evolving dynamics of societal opinions and sentiments. With the freedom of expression privilege granted after the Tunisian revolution, sensitive "taboo" topics such as political issues have become popular and widely discussed by Tunisians across social media platforms. However, on the downside, it has become easy to spread abusive/hate propaganda against individuals or groups. To address this gap, we introduce PoliTun, a new dataset designed specifically for political analysis in the Tunisian dialect, aiming to illuminate Tunisia's political landscape from a linguistic perspective. We describe the methodology used for constructing the PoliTun dataset, including data collection, preprocessing, and annotation. Then, we present experiments conducted with PoliTun for category detection and political opinion identification, utilizing various machine learning, deep learning and transformer-based models. The results reveal variable model performances. In conclusion, the development of PoliTun represents a significant advancement in political analysis in Tunisia, providing a foundation for nuanced exploration of political discourse in this unique sociopolitical landscape. PoliTun will be available upon request[1].

1 Introduction

Tunisia represents a country known for its vibrant political scene and diverse linguistic landscape, where the Tunisian Dialect represents its key form of expression. With the rise of social media platforms and the increasing prevalence of online discussions, the study of political discourse has gained prominence among researchers and policymakers. Particularly, the explosion of user-generated content on Twitter provides a valuable resource for mining insights on diverse subjects, including political content. On July, 25th, 2021, political decisions created a new political environment in Tunisia leading to various opinions across citizens. In fact, On July 25, 2021, the President of the Republic, Kaïs Saied, applied Article 80, dismissing the Head of Government, suspending all the activities of the National Assembly, and seizing full powers.

This situation split Tunisians into groups: loyalists to ideas with uncritical support, hunters for opportunities around, skeptics about choices, and opponents of slides towards an undemocratic political system. This cleavage caused intense debates between Tunisians, most of which took place on social media networks, leading to a high emergence of abusive/hate speech, polarization, and conspiracism. Hence, the need to create automatic solutions to detect such behaviours. However, the majority of existing studies tend to overlook the importance of regional dialects, thus limiting our understanding of political dynamics in specific linguistic contexts. The Tunisian Dialect, a variety of Arabic influenced by Berber, French, and other languages, serves as a distinct mode of communication among Tunisians, particularly in informal settings. It has unique linguistic features, idioms, and expressions that reflect the local cultural, historical, and political nuances. Therefore, an analysis of political discourse solely based on the Modern Standard Arabic or major languages would fail to capture the subtleties and intricacies embedded within the Tunisian Dialect. To address this research gap, we present the PoliTun dataset, a comprehensive collection of political texts in the Tunisian Dialect

[1]Please contact the authors via email to obtain the link.

sourced from Twitter. The dataset includes a wide range of political topics, covering elections, governance, public policies, activism, and more, to provide a holistic view of the political landscape in Tunisia. By focusing on the Tunisian Dialect, PoliTun enables researchers and language enthusiasts to go into the characteristics of political discourse within the unique socio-cultural context of Tunisia. In this paper, we provide an overview of related work in the field of political analysis and dialectal studies. Then, we describe the methodology employed in constructing the PoliTun data where we outline the methodology employed in compiling the PoliTun dataset, including data collection, preprocessing, and annotation procedure. Then, we present experiments performed using our dataset for category detection and political opinion detection. Finally, we present the conclusion and future work.

2 Related works

According to the Larousse online dictionary[2], an opinion can be defined as a judgment or feeling expressed by an individual or group on a subject or facts, reflecting their perception and thoughts, or a formal statement giving the reasons behind a given judgment. Opinion detection has become one of the most active fields of study in natural language processing since the early 2000s (Liu et al., 2010); (Liu, 2022). (Pang et al., 2008) conducted an in-depth study covering various aspects of opinion analysis, such as opinion extraction, sentiment classification, polarity analysis and opinion synthesis. This analysis included different approaches, whether lexicon-based or supervised learning, using similarity measures and various classifiers. Other work in the literature has compared several approaches, including naive bayes classifiers, support vector machines (SVMs) (Palau and Moens, 2009), logistic regression (Levy et al., 2014) . Applying the CNN deep learning approach in a mobile environment a study (Kalaivani and Jayalakshmi, 2021) proposed sentiment analysis based on movie reviews. Their approach involved the use of Polarity, IMDb and Rotten Tomato datasets, Results indicate that the integration of GloVe word vectors led to better performance. Recurrent and recursive neural networks were examined with different types of Arabic-specific processing (Al Sallab et al., 2015); (Al-Sallab et al., 2017); (Baly et al., 2017). Convolutional neural networks (CNNs) were trained using pre-trained word embeddings (Dahou et al., 2019). A hybrid model was proposed by (Farha and Magdy, 2019), where CNNs were used for feature extraction and LSTMs were used for sequence and context understanding. There is a lack of pre-trained language models, which limits the performance of NLP applications for some languages. The article (Kenton and Toutanova, 2019) highlights the ineffectiveness of traditional models as they require task-specific datasets, making them impractical to use. Unlike English, Arabic has a rich morphology and limited resources, as it has many dialects, which makes Automatic Comprehension of this language complex due to linguistic variations. There is a need to evaluate these models consistently and on various NLP tasks in Arabic. The article (Abdul-Mageed et al., 2020) presents the creation of two powerful language models specific to MarBERT and MarBERT-v2, pre-trained on massive and diverse datasets, including datasets, including social media data. Faced with these challenges, the researchers (Antoun et al., 2020) set out to create a solution capable of efficiently processing the Arabic language, it became imperative to develop a specific language processing model. Although pre-trained language models such as BERT have proven effective in English, their direct application to Arabic proved less conclusive. The morphological complexity of the Arabic language requires careful adaptation to exploit the full potential of these models. Hence, the idea of creating AraBERT emerged with the aim of meeting the unique challenges of the Arabic language and providing a powerful tool for the Arabic NLP community.

In (Abd et al., 2020), the authors present a Political Arabic Articles Data Set entitled PAAD, comprising 206 articles classified into Reform, Conservative, and Revolutionary collected from newspapers, social networks, general forums, and ideological websites. This data set is oriented at Arabic Computational Linguistics by providing a valuable resource for political text classification in Modern Standard Arabic and includes only three labels.

Despite this advancement, it does not cover the aspect of dialectal variations, which continues to be a significant challenge in Arabic NLP. To the best of our knowledge, our work presents the first dataset dedicated for political text in an Arabic dialect, particularly the Tunisian one.

[2]https://www.larousse.fr/

3 PoliTun Dataset: Data Building From Twitter

In this section, we detail the creation and the particularities of PoliTun, our large political opinion and category detection dataset composed of about 30K tweets collected about events after July 25th, 2021 Tunisian events.

The majority of the sentences deal with the post-July 25 period and the agitation of Tunisians on social networks in relation to the current political dynamic. In this work, we aim to offer a deep understanding of the diverse themes within political discussions, particularly, in the Tunisian context.

3.1 Data Collection

We collected tweets using the Twitter streaming API. We have collected and scraped tweets using Twitter hashtags related to the Tunisian Political Context. We manually extracted a list of more than **200 hashtags** that were used for scrapping tweets. Examples are the following: #لا_للمحاكمات_العسكرية, #تفكيك_منظومة_الفساد_السياسي, #تونس_جمهورية_مدنية_ديمقراطية, etc.

In order to make sure that all tweets are after July 25th event, a python script was created to extract only tweets subsequent to that date. Another python script was created to split data into Arabic, Tunizi which is the Tunisian Dialect written in Latin letters, French and English sentences. Examples from the dataset and their translation are presented in Table 1.

3.2 Data Preprocessing

The collected tweets were not clean as they included many punctuation marks including, hashtags, Emojis, and more. Therefore, the following steps have been followed to ensure the quality of this dataset:

- **Removing Punctuation:**

  The tweets in the dataset contained a mix of Arabic and non-Arabic punctuation. in Tunisia, the use of punctuation in written texts is relatively low compared to other languages. Hence, these marks are generally outliers and do not contribute to the overall understanding of the text. Hence, different punctuation marks were removed using a python script.

- **Removing Emojis:** In Tunisia, the use of sarcasm is quite often, both in real life and over social networks which, generally, leads to carrying different meanings for the same emoji which can lead to potential confusion for the learning models. Since there are no available libraries that can translate emojis into Tunisian, we removed emojis present in our dataset.

- **Removing duplicates and retweets:** Duplicate data can skew the results of the analysis of results. Since the model will give it undue weight, it would lead to biased or inaccurate predictions.

- **Keeping only Arabic-letters tweets for annotation:** We kept only the tweets in Arabic letters for annotation. We leave including Latin letters as future work.

Finally, we sampled about 30,000 tweets written in Arabic letters for annotation in total.

3.3 Data Annotation

The labels of the dataset were identified by three sociologists, experts in their domain after reading multiple randomly selected examples of the scrapped tweets.

The dataset was annotated by three Tunisian female native speakers, that are involved in civil society, all aged 25. Due to the limited number of annotators, only 30k data in Arabic letters was annotated divided equally on the annotators.

To ensure the quality of the annotaion, we take 100 internally annotated examples and ask for review from the sociologists. If an annotator mislabels more than 25% of these examples, we discard the annotations and ask her to relabel them based on the comments of the sociologists.

In our dataset, each sentence is labeled twice, considering two distinct aspects: **category** and **opinion**. The category aspect includes six labels, while the opinion aspect consists of three labels.

3.3.1 Category Identification

Regarding the category label, Six categories were identified:

- **Hate Speech**: A tweet that discriminates, stigmatizes, or incites violence or prejudice against individuals or groups based on attributes such as race, ethnicity, religion, gender, sexual orientation, disability, or other characteristics.

Table 1: Political Data Examples.

Tweet	English Translation
لله يبارك تونس باهية وانت سبب بهاها	God bless you. Tunisia is nice and you're the reason for it.
تونس الجزائر بلد حدودي معانا و بينا مصالح	Tunisia, Algeria is a border country with us and we have interests together.

- **Polarization**: mainly expresses the position of being for or against a political position. It often conveys a strong and uncompromising stance, contributing to an "us versus them" mentality. Such tweets may use language that reinforces the separation between conflicting perspectives and may lack nuance or a willingness to engage in constructive dialogue.
- **Conspiracy**: expressed more in the words of a political person or his/her supporters. It expresses a belief or suspicion about a secretive or covert plan that is allegedly being carried out by powerful individuals or organizations.
- **Denunciation**: is the recounting of a political event without taking a position. It refers to the act of publicly condemning or expressing disapproval of someone or something, often due to perceived wrongdoing, unethical behavior, or actions contrary to societal norms or values. It involves making an official or public statement declaring strong disapproval or condemnation.
- **Skepticism**: an attitude of doubt or disbelief towards claims, beliefs, or assertions, particularly those that are commonly accepted or taken for granted. Skepticism can manifest in various forms, ranging from questioning the validity of specific claims to adopting a general stance of skepticism towards all knowledge claims until sufficient evidence is provided.
- **Off-Topic**: the case where the tweet doesn't deal with any political discourse or subject relating to the Tunisian context.

3.3.2 Opinion Identification

The same tweet is also annotated with one of the three opinion labels:

- **Positive**: expresses a positive opinion regarding an event/idea.
- **Negative**: expresses a negative opinion regarding an event/idea.
- **Neutral**: without interest/perspective in the subject.

Tweets that are annotated as off-topic do not have an opinion label.

Examples from the dataset with their annotations are presented in Table 2.

3.4 Data Statistics

Statistics of the initial data after preprocessing and cleaning are presented in Table 3.

The annotated data includes about 30k tweets written in Arabic letters.

Table 4 and Figure 1 show the distribution of the annotated tweets by category (Off-topic, Denunciation, Polarization, Conspiracy, Skepticism, Hate speech). We note a clear superiority in number of comments categorized as "Off Topic", with a total of 12,503 annotations representing 42% of the data, compared to the other categories. The "Denunciation" and "Polarization" categories come next, with 7,763 and 7,660 annotated comments representing 26% and 25% respectively. "Conspiracy" includes 1,072 annotations, "Skepticism" with 528 annotations, and finally "Hate speech" with 448 annotations. The three last labels present only 7% of the dataset.

Figure 2 summarizes the distribution of the annotated tweets according to their opinion polarity (Positive, Negative, Neutral). We note that Negative comments outnumber Positive ones by a large margin: 57% were annotated as negative, 32% as positive, and 11% as neutral.

4 Experimental Setup

We divide our data into 80% for training and 20% for testing. We run experiments using the following machine learning and deep learning models: Naive Bayes (NB), Support Vector Machines (SVM), Logistic Regression (LR), Convolutional

Table 2: Political Data Examples with their annotation.

Tweet	English Translation	Category	Opinion
ياخي وينهم مواطنون ضد الإنقلاب أكاهوا خرجو مرتين ولاثلاثة	where are the Citizens that are against the Coup, they've been out two or three times	Polarization	Negative
والا ترهدين الطبوبي	The hypocrisy of Taboubi	Hate Speech	Negative

Language	#Sentences
Arabic	113341
Tunizi	3685
French	270
English	657
Total	**117953**

Table 3: Initial Data statistics.

Category	#Sentences
Off-topic	12503
Denunciation	7763
Polarization	7660
Conspiracy	1072
Skepticism	528
Hateful speech	448
Total	**29974**

Table 4: Distribution of Annotated Comments by Category.

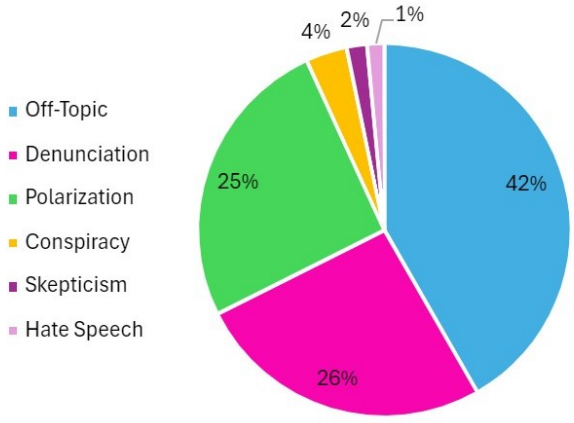

Figure 1: Category Label Percentages.

Neural Networks (CNN), and Long-Short Term Memory (LSTM).

When training deep learning models, we use as embedding FastText pretrained on Arabic Wikipedia, batch size equal to 128, and 20 epochs.

Also, different pre-trained models were used in order to achieve the best results. Bacause there is a

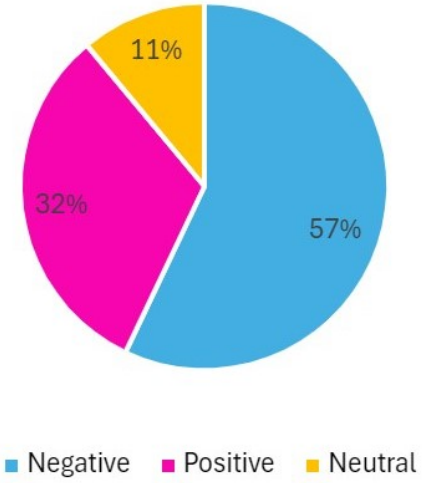

Figure 2: Opinion Label Percentages.

lack of pretained models for the Tunisian dialect, we chose to experiment with the following models that are pretrained on Arabic language and Arabic dialects respectively:

- **AraBERT** (Antoun et al., 2020): is a BERT based model for Modern Standard Arabic Language understanding, trained on 70M sentences from several public Arabic datasets and news websites.
- **MarBERT** (Abdul-Mageed et al., 2020): is a large-scale pretrained language model using the BERT base's architecture. MARBERT is trained on on 128 GB of tweets from various Arabic dialects containing at least 3 Arabic words. With very light preprocessing the tweets were almost kept at their initial state to retain a faithful representation of the naturally occurring text.

We finetuned BERT models with the following hyperparameters: batch size equal to 128, 20 epochs, and max seq length of 128.

The metrics used to evaluate the model's predictions are accuracy, recall, precision and F1 score

(micro and macro).

5 Results and Discussion

In this section, we present the set of experiments performed on both tasks and discuss the results obtained.

5.1 Opinion Classification subtask

Table 5 presents results of the category classification task with Kmeans, Logistic Regression, Naive Bayes and Support Vector Machines.

The results of the evaluation of machine learning models reveal varying metrics analyzed. The best result was acheived using SVM performing 70% accuracy. However, F1 macro score achieved 57% because the dataset is not balanced.

Model	LR	NB	SVM
Accuracy	0.69	0.68	**0.70**
F1 micro	0.69	0.68	0.70
F1 macro	0.58	0.42	0.57
Recall	0.62	0.68	0.55
Precision	0.57	0.76	0.67

Table 5: Performance Metrics of Machine Learning Models for Opinion Classification.

Table 6 presents the results of the deep learning algorithms for the opinion classification task. The CNN outperforms the LSTM model reaching an accuracy value of 95% and F1 macro of 91%.

Model	CNN	LSTM
Accuracy	**0.95**	0.65
F1 micro	0.92	0.68
F1 macro	0.91	0.57
Recall	0.91	0.53
Precision	0.95	0.62

Table 6: Performance Metrics of Deep Learning Models for Opinion Classification.

Table 7 presents results for finetuning the AraBERT and MarBERT models on the Opinion Classification task. In fact, AraBERT achieves an accuracy of 75.42%. while MarBERT outpeforms it acheiving 76% of accuracy measure and 63% F1 macro. This is mainly because our dataset is written in the Tunisian dialect and MarBERT was trained on different dialectal Arabic texts while AraBERT was trained on Modern Standard Arabic (MSA).

Model	AraBERT	MarBERT
Accuracy	0.75	**0.76**
F1 micro	0.64	0.65
F1 macro	0.62	0.63
Recall	0.62	0.63
Precision	0.67	0.66

Table 7: Performance Metrics of BERT variant Models for Opinion Classification.

5.2 Category Classification subtask

Table 8 presents results of the category classification task with Kmeans, Logistic Regression, Naive Bayes, and Support Vector Machines.

Model	Kmeans	LR	NB	SVM
Accuracy	0.45	0.64	0.60	**0.65**
F1-micro	0.45	0.64	0.60	0.65
F1-macro	0.24	0.37	0.21	0.33
Recall	0.30	0.34	0.21	0.31
Precision	0.34	0.45	0.54	0.64

Table 8: Performance Metrics of Machine Learning Models for Category Classification.

In this task, SVM also outperforms the other machine learning models achieving 65% accuracy and 33% F1 macro. In this subtask, F1 macro gives low results because we have a non balanced dataset.

Table 9 presents results of the category classification task with Convolutional Neural Networks and Long Short Term Memory.

Model	CNN	LSTM
Accuracy	**0.70**	0.54
F1 micro	0.96	0.59
F1 macro	0.35	0.33
Recall	0.35	0.29
Precision	0.36	0.64

Table 9: Performance Metrics of Deep Learning Models for Category Classification.

The CNN outperforms LSTM by achieving 70% and 35% accuracy and F1 macro results respectively. Again, due to the imbalance labels in the dataset, F1 macro achieves low results.

Table 10 presents results of finetuning AraBERT and MarBERT models for category identification task.

In this case, MarBERT outperforms AraBERT in terms of accuracy by 3%. However, AraBERT outperforms MarBERT it terms of F1 macro by

Model	AraBERT	MarBERT
Accuracy	0.66	**0.69**
F1 micro	0.32	0.35
F1 macro	0.39	0.32
Recall	0.30	0.32
Precision	0.34	0.37

Table 10: Performance Metrics of BERT variant Models for Category Classification.

7%. But, still achieving low results because we are dealing with an imbalance in the dataset labels.

6 Conclusion and Future work

The development of PoliTun dataset represents a significant step forward in political analysis, particularly within the context of the Tunisian dialect. By creating a dataset of about 30,000 manually annotated data by a team work of both Tunisian sociologists and engaged citizens, we tackled the science of political discourse in Tunisia. Our experiments have showcased the potential of computational methods in understanding complex linguistic dynamics within this unique sociopolitical landscape. Moving forward, several avenues for future research present themselves. Firstly, expanding the PoliTun dataset to include larger range of topics and dialectical nuances which would enhance its utility for comprehensive political analysis. Also, including collaboration between researchers, policymakers, and local communities in Tunisia to co-create and utilize PoliTun for informed decision-making and civic engagement initiatives would be instrumental in maximizing its societal impact. Overall, the continued development and utilization of PoliTun stand to enrich our understanding of political dynamics in Tunisia and beyond, contributing to more inclusive and data-driven approaches to governance and social change.

References

Dhafar Hamed Abd, Ahmed T Sadiq, and Ayad R Abbas. 2020. Paad: Political arabic articles dataset for automatic text categorization. *Iraqi Journal for Computers and Informatics*, 46(1):1–11.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020. Arbert & marbert: Deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*.

Ahmad Al-Sallab, Ramy Baly, Hazem Hajj, Khaled Bashir Shaban, Wassim El-Hajj, and Gilbert Badaro. 2017. Aroma: A recursive deep learning model for opinion mining in arabic as a low resource language. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 16(4):1–20.

Ahmad Al Sallab, Hazem Hajj, Gilbert Badaro, Ramy Baly, Wassim El-Hajj, and Khaled Shaban. 2015. Deep learning models for sentiment analysis in arabic. In *Proceedings of the second workshop on Arabic natural language processing*, pages 9–17.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Ramy Baly, Hazem Hajj, Nizar Habash, Khaled Bashir Shaban, and Wassim El-Hajj. 2017. A sentiment treebank and morphologically enriched recursive deep models for effective sentiment analysis in arabic. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 16(4):1–21.

Abdelghani Dahou, Mohamed Abd Elaziz, Junwei Zhou, and Shengwu Xiong. 2019. Arabic sentiment classification using convolutional neural network and differential evolution algorithm. *Computational intelligence and neuroscience*, 2019.

Ibrahim Abu Farha and Walid Magdy. 2019. Mazajak: An online arabic sentiment analyser. In *Proceedings of the fourth arabic natural language processing workshop*, pages 192–198.

MS Kalaivani and S Jayalakshmi. 2021. Sentiment analysis on micro-blog data using machine learning techniques-a review. In *IOP Conference Series: Materials Science and Engineering*, volume 1049, page 012012. IOP Publishing.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500.

Bing Liu. 2022. *Sentiment analysis and opinion mining*. Springer Nature.

Bing Liu et al. 2010. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2(2010):627–666.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135.