TUNIPROD: An Optimized ProdLDA-Based Multi-Stage Framework for Cross-Lingual Topic Modeling of Code-Switched Public Discourse

1st Azer Mahjoubi
RLANTIS Laboratory
University of Tunis El Manar
Tunis, Tunisia
0000-0002-6096-6048

2nd Samawel Jaballi
Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI,
UMR 5800, F-33400 Talence, France
0009-0002-5476-9539

3rd Mounir Zrigui
RLANTIS Laboratory
University of Monastir
Monastir, Tunisia
0000-0002-4199-8925

***Abstract*—The rapid proliferation of social data during global crises offers unprecedented opportunities for real-time monitoring; yet, existing research remains constrained by a monolingual bias, often failing in linguistically diverse regions characterized by complex code-switching. This paper addresses this gap by introducing TUNIPROD, an optimized, multi-stage framework designed for cross-lingual topic discovery within the Tunisian digital landscape—a unique environment where Arabic, French, English, and Arabizi (Latin-scripted Arabic) intersect. We present a robust data processing pipeline that aggregates 177,000 documents from diverse sources, including social media, TV, and radio, and employs a multi-stage normalization process specifically tailored for Arabizi. Central to the TUNIPROD framework is the optimization of the Product of Experts Latent Dirichlet Allocation (ProdLDA) model through a systematic grid search of Dirichlet hyperparameters and expert instances. Our results demonstrate that the optimized TUNIPROD framework achieves a high topic coherence score of 0.89, significantly outperforming standard LDA models in capturing the nuanced thematic dependencies of crisis-related discourse. We identify seven dominant macro-topics, revealing that public discourse was primarily driven by Economic (23%) and Health (22%) concerns, with distinct thematic emergence for Vaccination and Governance.**

***Index Terms*—Topic modeling, ProdLDA, Code-Switching, Cross-lingual, Social Media Analysis, Tunisian Dialect, multilingual NLP**

I. Introduction

The COVID-19 pandemic set off an unprecedented global "infodemic," with social media platforms becoming the central sites for public discourse, information dissemination, and the formation of public opinions. However, the inherent multilingualism of these platforms presents a major challenge for conventional analytical methods, especially in linguistically diverse regions like Tunisia [1]. The Tunisian digital landscape, defined by the fluid use of Arabic, French, English, and Arabizi (a form of code-switching using a Latin-based script for Arabic), creates a fundamental problem for extracting latent thematic trends via standard monolingual models.

This research resolves this problem by employing the Product of Experts Latent Dirichlet Allocation (ProdLDA) model, optimized with a grid search strategy, to tackle these multilingual complexities. By moving beyond linguistic barriers, our work aims to reveal meaningful patterns and recurrent themes, contributing to both computational linguistics and the sociocultural analysis of societies during global crises. Specifically, the core contributions of this paper are:

1) A comprehensive analysis of a novel, large-scale multilingual dataset harvested from diverse Tunisian media sources.
2) The successful application and evaluation of the TUNIPROD framework to perform topic discovery in a complex code-switching environment.
3) Actionable insights into the thematic evolution of public discourse in Tunisia during the pandemic, offering a methodological precedent for similar multilingual studies.

The remainder of this paper is structured as follows. Section II reviews the relevant literature. Section III details our data collection and modeling approach. Section IV presents the experimental results. Section V concludes the paper and outlines future work.

II. Background and Related Work

This research builds upon two primary streams of work: the use of social media for crisis monitoring and the application of topic modeling for textual analysis.

A. Social Media as a Real-Time Societal Sensor

Social media platforms have been widely recognized as invaluable real-time sensors for monitoring public sentiment and disseminating information during large-scale events. In the context of natural disasters, early studies demonstrated the efficacy of microblogging platforms. For instance, Sakaki et al. [4] analyzed over 1.5 million tweets following the 2011 Great East Japan Earthquake, successfully monitoring awareness and anxiety levels in real-time. Similarly, Kadhem et al. [5] explored post-earthquake communication patterns, revealing how platforms facilitate direct communication within affected areas and broader information sharing via retweets.

This paradigm extends to public health crises. Platforms like Facebook and Reddit have become crucial data sources for

understanding public concerns during epidemics. Sharma et al. [7] analyzed Facebook data during the Zika virus outbreak to understand public reaction and information spread. These studies collectively affirm the role of social media as a digital mirror of public sentiment, indispensable for effective crisis response and information management [4], [19].

B. Topic Modeling for Public Discourse Analysis

The utility of topic modeling extends powerfully to the analysis of unsolicited patient narratives found in large-scale online communities. Platforms such as Reddit, with its pseudonymous nature and specialized forums (subreddits), provide a rich source of candid data on sensitive health topics that are often underreported in clinical settings. In the domain of maternal health, for example, studies [11], [18] have employed topic modeling to sift through tens of thousands of posts and comments. This computational lens has enabled the automatic discovery of a fine-grained topical structure within these conversations, pinpointing key concerns ranging from prenatal anxieties and medication safety to more sensitive issues like postpartum depression and social isolation [3]. The value of such findings is twofold: they offer public health officials an authentic, real-time overview of population-level concerns, and they highlight specific gaps in patient support systems that can be addressed with targeted interventions and resources.

More recently, this technique proved vital for analyzing discourse surrounding the COVID-19 pandemic. Moghadasi and Mohammadzadeh [8] utilized topic modeling to discern prevalent themes in Persian scientific articles about COVID-19. Their analysis uncovered a wide spectrum of topics, from clinical aspects and prevention strategies to the societal and psychological repercussions of the pandemic [2]. However, a limitation of many existing studies is their focus on monolingual corpora. The challenge intensifies in regions characterized by code-switching and multilingualism, where standard models fail to capture the full semantic richness of the discourse. Our work directly confronts this challenge by applying an advanced topic model to the linguistically complex Tunisian context.

III. Methodology

Our methodology is structured as a comprehensive pipeline, encompassing corpus creation, data preprocessing and normalization, and multilingual topic modeling. This section details the first two stages of this pipeline.

A. The Multilingual Tunisian COVID-19 Corpus

The foundation of our analysis is a large-scale, multilingual corpus specifically designed to capture a comprehensive perspective of the Tunisian public discourse during the COVID-19 pandemic. To ensure ecological validity, the corpus integrates data from four distinct sources, balancing formal and informal communication styles:

- **ASTD** [10]: Formal texts in Modern Standard Arabic (MSA) from official news and governmental sources.
- **CSTA** [11]: User-generated content from social media, primarily in Tunisian dialectal Arabic and Arabizi.
- **TUNIZI** [22]: A diverse corpus containing Arabizi and Latin-script languages (French, English) from YouTube and Twitter.
- **TSAC** [23]: Comments from Tunisian TV and radio Facebook pages, capturing public reactions to media broadcasts.

By merging these corpora, we compiled a substantial dataset of 177,000 documents, reflecting the complex linguistic landscape of Tunisia. Table I provides an overview of the corpus statistics, while Table II details its linguistic composition, highlighting the significant presence of code-switching phenomena.

TABLE I
Overall Statistics of the Aggregated Corpus

Docs	Sentences	Tokens	Vocab
177,000	259,031	5,107,110	114,779

TABLE II
Approximate Language Distribution Within the Corpus

% Arabic	% French	% Arabizi[a]	% English
39	19	23	19

[a] Code-switching / transliterated Arabic.

B. Data Preprocessing and Normalization

To prepare the raw, noisy text for topic modeling, we implemented a multi-stage preprocessing pipeline, visualized in Fig. 1. This pipeline consists of two primary phases: text cleansing and language normalization.

1) Text Cleansing: Following a methodology inspired by Mulki et al. [14], we applied a series of standard and language-specific cleansing steps. The primary objective was to reduce feature sparsity and eliminate noise that could interfere with the topic discovery process. These steps included:

- **Removal of non-essential elements**: Punctuation, special characters, URLs, and numeric digits were removed from the text.
- **Arabizi-specific normalization**: We substituted numerals commonly used in Arabizi with their corresponding Arabic letters (e.g., "3" $\rightarrow$ Ayn, "5" $\rightarrow$ Kha). This step helps consolidate vocabulary.
- **Tokenization**: The multilingual text was segmented into individual tokens.
- **Stopword removal**: A custom, consolidated list of stopwords for Arabic, French, and English was used to filter out high-frequency, low-information words.
- **Stemming**: We employed a light stemmer to reduce words to their root form (e.g., "vaccination" $\rightarrow$ "vaccin"), further unifying the vocabulary.

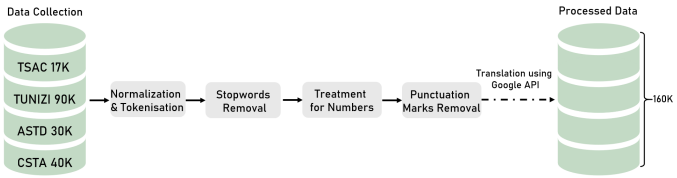

Fig. 1. Overview of the Data Preprocessing Pipeline

2) Language Normalization via Translation: A core challenge of our dataset is its inherent multilingualism. To apply a unified topic model, we chose to normalize the entire corpus into a single target language: English. This approach, while potentially losing some cultural nuance, creates a homogeneous linguistic space where topics can be modeled consistently across the entire dataset. For this task, we leveraged the `Google Translate API`, a tool recognized for its high accuracy and efficiency in machine translation. This step is critical for homogenizing the feature space, ensuring that a concept expressed in Arabic, French, or Arabizi is mapped to a common representation in English. The quality of this translation step is crucial and was systematically evaluated using standard metrics, as detailed in our experiments section (Section IV).

C. Topic Modeling Approach

Once preprocessing and normalization are complete, we extract latent thematic structures from the homogenized corpus using a probabilistic topic modeling approach designed to handle the complexity and scale of our dataset, depicted in Fig. 2. Our approach begins with Rapid Automatic Keyphrase Extraction (`RAKE+`) to identify dominant terminology within the corpus, providing a qualitative foundation for our model configuration.

This initial analysis suggested the presence of approximately seven distinct macro-topics related to the pandemic discourse (e.g., economy, health, governance). Consequently, we set the number of topics K to 7 for the primary modeling phase. The cornerstone of this phase is the ProdLDA model [15], an extension of standard LDA particularly well-suited for capturing the interconnected nature of public discourse during a crisis. ProdLDA is a generative probabilistic model that assumes each document is a mixture of various topics, where each topic is characterized by a distribution of words. Its key innovation lies in the use of an "experts" mechanism, which models dependencies between topics by representing them as a product of simpler distributions [16]. This feature is especially valuable in our context, since themes like "economic impact" are inherently related. To formalize this intuition, we provide a summary of the model's variables in Table III.

The generative process is defined as follows: for each document d, a topic distribution θ_d is drawn from Dirichlet(α); for each word w_{dn}, a topic z_{dn} is sampled from θ_d, and the word is drawn from the corresponding topic distribution $\phi_{z_{dn}}$:

$$\theta_j \sim \text{Dir}(\alpha), \quad \phi_k \sim \text{Dir}(\beta), \quad z_{ij} \sim \theta_j, \quad x_{ij} \sim \phi_{z_{ij}} \quad (1)$$

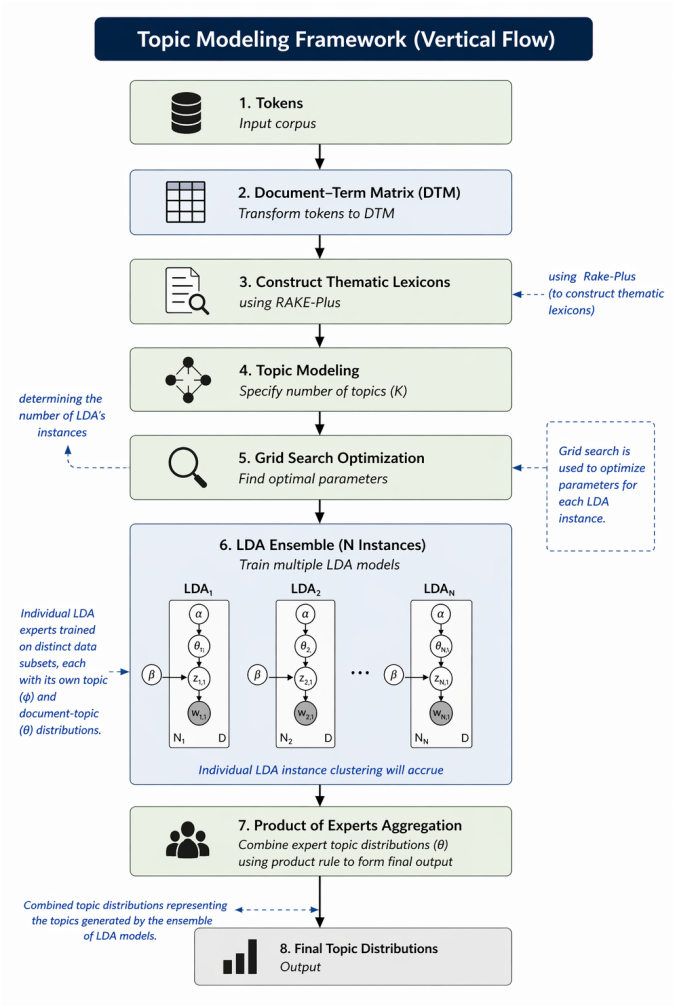

Fig. 2. Overview of the proposed Topic Modeling Strategy

TABLE III
SUMMARY OF VARIABLES USED IN THE PRODLDA MODEL

Variable	Role
D	Number of documents
V	Vocabulary size (unique words)
N_d	Number of words in document d
K	Number of latent topics
w_{dn}	The n-th word in document d
z_{dn}	Topic assigned to w_{dn}
θ_d	Per-document topic proportions
ϕ_k	Per-topic word distribution
α, β	Dirichlet priors for θ_d and ϕ_k

The full joint probability distribution is:

$$p(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi} \mid \alpha, \beta) = \prod_{d=1}^{D} p(\theta_d \mid \alpha) \prod_{k=1}^{K} p(\phi_k \mid \beta) \cdot \prod_{n=1}^{N_d} p(z_{dn} \mid \theta_d)\, p(w_{dn} \mid z_{dn}, \boldsymbol{\phi}) \quad (2)$$

Given observed words $\mathbf{w}$, the central task of inference is to compute the posterior distribution of the latent variables (z, θ, ϕ). This is typically achieved using approximation methods such as Variational Inference or Gibbs Sampling.

Following this theoretical framework, we designed a rigorous experimental protocol to apply the ProdLDA model to our preprocessed dataset. Using insights from our initial keyword extraction, we established $K = 7$ macro-categories: *economy, governance, sport, health, education, vaccination,* and *others*. The core challenge then lies in optimizing the ProdLDA model to ensure that the discovered topics are both coherent and meaningful. The performance of a topic model is highly sensitive to its hyperparameters, primarily the Dirichlet priors α (which governs the sparsity of per-document topic distributions) and β (which controls the sparsity of per-topic word distributions).

To find the optimal configuration for our specific dataset, we implemented a systematic grid search methodology. This process involves training multiple instances of the ProdLDA model while varying key hyperparameters, including α, β, and the number of model "experts" (LDA instances), up to a maximum of 10 for computational efficiency. The effectiveness of each configuration is quantitatively measured using a topic coherence score, a standard metric that correlates well with human judgments of topic interpretability. This systematic tuning is vital for balancing model complexity with semantic coherence, ensuring that our final model is well-suited to the nuanced dynamics of the multilingual dataset. The specific results of this optimization process, along with our evaluation of the initial translation quality and the final analysis of the discovered topics, are detailed in the subsequent section.

IV. Experiments and Results

A. Translation Quality Assessment

A crucial first step for our topic modeling is the successful normalization of the multilingual corpus into a single target language, English. To validate the reliability of this step, we evaluated the performance of the `Google Translate API` across each of the four source corpora. We used two standard metrics from the field of machine translation:

- **BLEU** (Bilingual Evaluation Understudy) Score: Measures the n-gram precision between the machine-translated text and a set of high-quality human references, providing an indicator of fluency and adequacy. Higher scores are better.
- **TER** (Translation Edit Rate): Calculates the minimum number of edits (insertions, deletions, substitutions) required to transform the machine-translated text into a human reference. It serves as an error metric, where lower scores indicate higher accuracy.

The evaluation results are summarized in Table IV. The table presents the average BLEU and TER scores for each sub-corpus, weighted by the number of samples.

TABLE IV
Translation Quality Metrics per Source Dataset

Dataset	Language Focus	N	BLEU	TER
TUNIZI	Arabizi, Fr., En.	90k	0.90	0.08
ASTD	Modern Standard Ar.	30k	0.88	0.09
TSAC	Tunisian Dial., Arabizi	17k	0.86	0.11
CSTA	Tunisian Dial., Arabizi	40k	0.85	0.12

The results show strong translation quality across all datasets. The consistently high BLEU scores (ranging from 0.85 to 0.90) and low TER values (0.08 to 0.12) confirm that the translated text is both fluent and accurate, preserving the semantic intent of the original content. As expected, the `TUNIZI` corpus, which already contains a significant portion of Latin-script text, achieved the highest performance. In contrast, the corpora rich in colloquial Tunisian dialect (`CSTA` and `TSAC`) presented a greater challenge for the translation model, resulting in slightly lower scores. Nevertheless, the overall performance is robust, providing a reliable monolingual foundation for the subsequent topic modeling phase.

B. Topic Model Optimization and Analysis

Following the successful validation of our translation pipeline, we proceeded to the core task of topic discovery. To ensure the semantic quality and interpretability of the topics generated by the `ProdLDA` model, we conducted a systematic, two-stage grid search over its key hyperparameters. The performance of each model configuration was evaluated using the C_v coherence score, a standard metric that measures the semantic similarity of high-scoring words within a topic. Higher scores indicate more human-interpretable topics.

Stage 1 — Number of Experts. First, we investigated the impact of the number of ProdLDA "experts" (or instances) on topic coherence. Due to computational constraints, we explored configurations with up to 10 instances. The results, presented in Table V, reveal a clear trend. The coherence score improves significantly as the number of instances increases from one to seven, peaking at **0.89**. Beyond this point, adding more instances leads to a slight degradation in performance, suggesting a point of diminishing returns where model complexity begins to outweigh the benefits. Based on this finding, we fixed the optimal number of instances at seven for the subsequent optimization phase.

Stage 2 — Dirichlet Hyperparameters. With the number of instances set to seven and $K = 7$, we performed a second grid search to fine-tune the Dirichlet hyperparameters α and β (Table VI). The results show the model's sensitivity to these parameters. The highest coherence score of **0.88** was achieved with $\alpha = 0.1$ and $\beta = 0.3$. This configuration suggests that a moderately sparse topic distribution per document, combined

TABLE V
IMPACT OF NUMBER OF PRODLDA INSTANCES ON COHERENCE SCORE

Config.	Instances	Coherence
1	1	0.72
2	2	0.81
3	4	0.84
4	5	0.78
5	6	0.85
6	**7**	**0.89**
7	9	0.86

with a slightly denser word distribution per topic, yields the most interpretable thematic structure for our dataset.

TABLE VI
IMPACT OF DIRICHLET HYPERPARAMETERS ON COHERENCE SCORE (7 INSTANCES, $K = 7$)

α	β	Coherence
0.01	0.1	0.75
0.05	0.2	0.82
0.1	**0.3**	**0.88**
0.15	0.4	0.85
0.2	0.5	0.79
0.25	0.6	0.87
0.3	0.7	0.81

Having identified the optimal configuration (7 instances, $K = 7$, $\alpha = 0.1$, $\beta = 0.3$), we applied it to the full dataset. Fig. 3 presents a quantitative overview of the extracted topics.

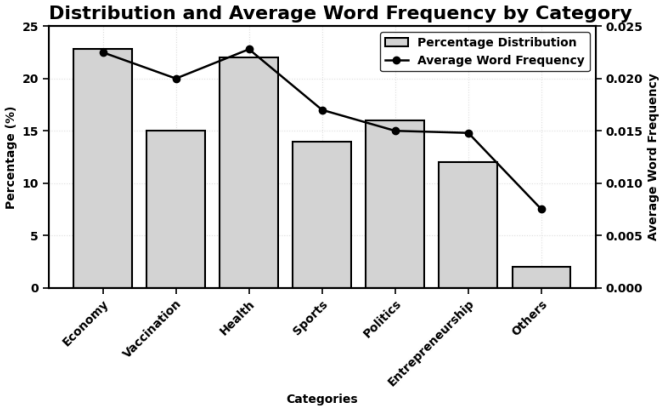

Fig. 3. Distribution and average word frequency of the seven discovered topics

The topic distribution reveals that the public discourse during the pandemic was overwhelmingly dominated by conversations surrounding the **Economy** (approx. 23%) and **Health** (22%). These two core themes collectively account for nearly half of the entire dataset, underscoring the dual public health and economic nature of the crisis. Other highly significant topics include **Politics** (16%) and **Vaccination** (15%), reflecting the intense public debates over government measures and the vaccine rollout. A deeper insight is offered by the average word frequency: topics like "Economy" and "Health" are constructed from more common, high-frequency terms, while the "Vaccination" topic, though prevalent, is defined by a more specialized and less frequent vocabulary (e.g., brand names like "Pfizer", technical terms like "dose").

C. Discussion

Our work demonstrates the successful application of the TUNIPROD framework to uncover multilingual topics within Tunisia's digital discourse during the COVID-19 pandemic. The analysis yielded clear insights into the public discourse and validated our methodological approach. The prominence of "Economy" (23%) and "Health" (22%) as main topics underscores their pervasive influence on public discussions, a pattern observed globally but uniquely captured here from a multilingual perspective. The distinct emergence of "Vaccination" as a primary topic, rather than a sub-theme of health, highlights its significant role in shaping public conversation. Furthermore, the strong presence of the "Politics" topic points to a substantial public focus on governmental crisis management, reflecting intense scrutiny of public trust and policy effectiveness.

Our findings robustly support the efficacy of the TUNIPROD framework's two-step method: initial language normalization followed by optimized topic modeling. The high translation quality scores, particularly for high-resource languages like French and Arabic, confirm the effectiveness of this strategy, consistent with prior research [20]. Crucially, the successful optimization of ProdLDA within TUNIPROD demonstrated superior performance in managing the subtle thematic dependencies inherent in social media data, extending beyond the capabilities typically achieved by standard LDA models in previous analyses [8].

However, we acknowledge several limitations and inherent risks within the TUNIPROD approach. The reliance on the Google Translate API for language normalization, while generally effective, may not fully capture the intricate cultural nuances, idiomatic expressions, or specific slang prevalent in the Tunisian dialect. This limitation could lead to a partial loss of semantic information, potentially affecting the granularity and precision of topic interpretation. Moreover, the current methodology is exclusively focused on topic identification and does not integrate sentiment analysis. Consequently, while topics like "Vaccination" are successfully identified, the model cannot differentiate between positive, negative, or neutral sentiments expressed within that discourse, which is vital for a comprehensive understanding of public opinion. Finally, the dataset, primarily sourced from social media and digital platforms, inherently carries a risk of sampling bias. These platforms tend to amplify the voices of urban, digitally-savvy populations, meaning the findings may not fully represent the diverse perspectives of the entire Tunisian populace, thereby potentially skewing the observed thematic distributions.

V. CONCLUSION

This paper introduced TUNIPROD, an optimized ProdLDA-based multi-stage framework for cross-lingual topic modeling,

successfully applied to diverse Tunisian media sources during the COVID-19 crisis. By effectively integrating a robust data processing pipeline with an optimized ProdLDA model, TUNIPROD enabled the extraction of seven clear and meaningful topics, thereby overcoming the significant challenges posed by code-switching and multilingual data. Our work underscores the critical importance of analyzing cross-lingual data to achieve a holistic understanding of public discourse in globally connected and locally distinct societies. TUNIPROD provides researchers and policymakers with flexible method for crisis informatics, offering a practical tool for navigating the complexities of the multilingual digital world.

While TUNIPROD demonstrates significant advancements, it is important to acknowledge certain inherent difficulties and risks. These include the potential for subtle semantic loss during machine translation, the current lack of integrated sentiment analysis, and the inherent sampling bias of social media data, which may not fully represent the entire population. Future work will focus on addressing these limitations by exploring native multilingual topic models to minimize semantic loss and integrating sentiment-aware analysis for a more nuanced understanding of public opinion.

Competing Interests

The authors have no competing interests to declare.

References

[1] E. Hkiri, S. Mallat, M. Zrigui, and M. Mars, "Constructing a lexicon of Arabic-English named entity using SMT and semantic linked data," *International Arab Journal of Information Technology (IAJIT)*, vol. 14, no. 6, 2017.

[2] E. Hkiri, S. Mallat, and M. Zrigui, "Arabic-english text translation leveraging hybrid NER," in *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation, PACLIC 2018, Cebu City, Philippines, November 16-18, 2017*, R. E. Roxas, Ed. The National University (Phillippines), 2017, pp. 124–131. [Online]. Available: https://aclanthology.org/Y17-1019/

[3] A. Mahmoud and M. Zrigui, "Distributional semantic model based on convolutional neural network for arabic textual similarity," *International Journal of Cognitive Informatics and Natural Intelligence*, vol. 14, no. 1, pp. 35–50, 2020.

[4] F. Toriumi, T. Sakaki, K. Shinoda, K. Kazama, S. Kurihara, and I. Noda, "Information sharing on Twitter during the 2011 catastrophic earthquake," in *Proceedings of the 22nd International Conference on World Wide Web*, 2013, pp. 1025–1028.

[5] R. F. Kadhem, S. Mallat, E. Hkiri, A. Joudah, A. M. Albarrak, and M. Zrigui, "Towards a hybrid document indexing approach for Arabic documentary retrieval system," in *Advances in Computational Collective Intelligence*. Springer, Cham, 2023, pp. 262–271.

[6] S. Jaballi, M. J. Hazar, S. Zrigui, A. Mahjoubi, H. Nicolas, and M. Zrigui, "Decoding multilingual topic dynamics and trend identification through ARIMA time series analysis on social networks: A novel data translation framework enhanced by LDA/HDP models," *Journal of Electrical and Computer Engineering*, vol. 2024, no. 1, p. 6669491, 2024.

[7] M. Sharma, K. Yadav, N. Yadav, and K. C. Ferdinand, "Zika virus pandemic—analysis of Facebook as a social media health information platform," *American Journal of Infection Control*, vol. 45, no. 3, pp. 301–302, 2017.

[8] M. Dehghani and F. Ebrahimi, "ParsBERT topic modeling of Persian scientific articles about COVID-19," *Informatics in Medicine Unlocked*, vol. 36, p. 101144, 2023.

[9] S. Jaballi, S. Zrigui, M. J. Hazar, H. Nicolas, and M. Zrigui, "Exploring unsupervised word representations models and neural networks for informal multilingual text against COVID-19 social media content," in *Asian Conference on Intelligent Information and Database Systems*. Springer, 2024, pp. 347–359.

[10] M. Nabil, M. Aly, and A. F. Atiya, "ASTD: Arabic sentiment tweets dataset," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015, pp. 2515–2519.

[11] S. Jaballi, S. Zrigui, M. A. Sghaier, D. Berchech, and M. Zrigui, "Sentiment analysis of Tunisian users on social networks: Overcoming the challenge of multilingual comments in the Tunisian dialect," in *International Conference on Computational Collective Intelligence (ICCCI)*, ser. Lecture Notes in Computer Science, vol. 13501. Springer, 2022, pp. 176–192.

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

[13] R. Bhatia, P. Garg, and R. Johari, "Corpus based Twitter sentiment analysis," *SSRN Electronic Journal*, 2018.

[14] H. Mulki, H. Haddad, C. B. Ali, and I. Babaoglu, "Tunisian dialect sentiment analysis: A natural language processing-based approach," *Computación y Sistemas*, vol. 22, no. 4, pp. 1223–1232, 2018.

[15] R. Kadhem, S. Mallat, and M. Zrigui, "Survey semantic Arabic language in deep learning based recommendation system," in *2023 International Conference on Innovations in Intelligent Systems and Applications (INISTA)*. IEEE, 2023, pp. 1–8.

[16] S. Gayed, S. Mallat, and M. Zrigui, "Exploring word embedding for Arabic sentiment analysis," in *Recent Challenges in Intelligent Information and Database Systems*. Springer, Singapore, 2022, pp. 92–101.

[17] A. Mahmoud and M. Zrigui, "Siamese AraBERT-LSTM model based approach for Arabic paraphrase detection," in *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, Manila, Philippines, 2022, pp. 545–553.

[18] S. Jaballi, M. J. Hazar, S. Zrigui, H. Nicolas, and M. Zrigui, "Deep bidirectional LSTM network learning-based sentiment analysis for Tunisian dialectical Facebook content during the spread of the coronavirus pandemic," *International Conference on Computational Collective Intelligence*, vol. 13501, pp. 96–109, 2023.

[19] S. Jaballi, S. Zrigui, H. Nicolas, and M. Zrigui, "Analyzing multilingual conversations during COVID-19: an imbalanced class-ensemble learning approach with reweighted AdaBoost-SVM for code-switched text classification," *Research Square*, 2024, preprint.

[20] S. Jaballi, M. J. Hazar, S. Zrigui, H. Nicolas, and M. Zrigui, "Resnet-based pandemic keyword spotting in continuous multilingual speech: A study in unesco's audio messages for rapid health response," in *International Conference on Computational Collective Intelligence*. Springer, 2025, pp. 76–91.

[21] M. J. Hazar, S. Jaballi, M. Maraoui, M. Zrigui, and H. Nicolas, "A hybrid e-learning recommendation system incorporating user reviews and ratings for enhanced course selection," 2025.

[22] C. Fourati, A. Messaoudi, and H. Haddad, "TUNIZI: a Tunisian Arabizi sentiment analysis dataset," *arXiv preprint arXiv:2004.14303*, 2020. [Online]. Available: https://arxiv.org/abs/2004.14303

[23] S. Medhaffar, F. Bougares, Y. Estève, and L. Hadrich-Belguith, "Sentiment analysis of Tunisian dialects: Linguistic resources and experiments," in *Proceedings of the Third Arabic Natural Language Processing Workshop*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 55–61. [Online]. Available: https://aclanthology.org/W17-1307/