The Nakba Lexicon: Building a Comprehensive Dataset from Palestinian Literature

Izza Abu Haija[1], Salim Almandhari[2], Mo El-Haj[2], Jonas Sibony[3], Paul Rayson[2]

[1]Freie Universität Berlin, Berlin, Germany
[2]UCREL NLP Group, Lancaster University, Lancaster, UK
[3]Department of Arabic and Hebrew Studies, Sorbonne Université, Paris, France - CERMOM

Correspondence: izza.abuhaija@gmail.com

Abstract

This paper introduces the Nakba Lexicon, a comprehensive dataset derived from the poetry collection *Asifa 'Ala al-Iz'aj* (Sorry for the Disturbance) by Istiqlal Eid, a Palestinian poet from El-Birweh. Eid's work poignantly reflects on themes of Palestinian identity, displacement, and resilience, serving as a resource for preserving linguistic and cultural heritage in the context of post-Nakba literature. The dataset is structured into ten thematic domains, including political terminology, memory and preservation, sensory and emotional lexicon, toponyms, nature, and external linguistic influences such as Hebrew, French, and English, thereby capturing the socio-political, emotional, and cultural dimensions of the Nakba. The Nakba Lexicon uniquely emphasises the contributions of women to Palestinian literary traditions, shedding light on often-overlooked narratives of resilience and cultural continuity. Advanced Natural Language Processing (NLP) techniques were employed to analyse the dataset, with fine-tuned pre-trained models such as ARABERT and MARBERT achieving F1-scores of 0.87 and 0.68 in language and lexical classification tasks, respectively, significantly outperforming traditional machine learning models. These results highlight the potential of domain-specific computational models to effectively analyse complex datasets, facilitating the preservation of marginalised voices. By bridging computational methods with cultural preservation, this study enhances the understanding of Palestinian linguistic heritage and contributes to broader efforts in documenting and analysing endangered narratives. The Nakba Lexicon paves the way for future interdisciplinary research, showcasing the role of NLP in addressing historical trauma, resilience, and cultural identity.

1 Introduction

The Nakba, meaning "catastrophe" in Arabic, marks a significant chapter in Palestinian history, signifying the mass displacement and loss of homeland that followed the establishment of the State of Israel in 1948. This event not only reshaped Palestinian society but also deeply affected its linguistic, cultural, and political landscapes. Palestinian literature has since played a vital role in documenting and preserving the memory, identity, and collective experiences of Palestinians across generations. Within this literary tradition, the works of poets such as Istiqlal Eid serve as essential records of the Palestinian narrative, capturing the nuanced emotional and socio-political complexities faced by Palestinian communities (Sa'di and Abu-Lughod, 2007).

Istiqlal Eid (?[1]), a Palestinian poet from El-Birweh[2], uses her poetry to convey a powerful perspective on identity, exile, and resistance. Her collection, *Asifa 'Ala al-Iz'aj* (Sorry for the Disturbance), weaves a tapestry of themes central to the Palestinian experience, from cultural memory to the impact of displacement. As a female poet writing in both Modern Standard Arabic and Palestinian dialect, Eid's work provides a unique vantage point into the resilience of Palestinian identity, amplified by her familial connection to the iconic poet Mahmoud Darwish. This connection grounds her in the cultural and historical heritage of her homeland while navigating the challenges of being a refugee within her own country (Eid, 2017).

This study builds on Eid's poetry to cre-

[1]The author refuses to disclose her birthdate or age, and thus it is represented as "?" in respect of her wishes

[2]El-Birweh, also spelled as Al-Birwa, was a Palestinian village, located 10.5 kilometres (6.5 miles) east of Acre (Akka). The village was depopulated during the 1948 Arab-Israeli conflict and subsequent wars.

Proceedings of the Proceedings of the 1st International Workshop on Nakba Narratives as Language Resources, pages 37–47
January 20, 2025.

ate a Nakba Lexicon, a structured dataset of terms relevant to the Palestinian experience post-1948. By categorising terms from Eid's poetry into thematic domains—such as political terminology, memory and preservation, sensory and emotional lexicon, toponyms, and Hebrew linguistic influence—we aim to provide a comprehensive resource for Natural Language Processing (NLP) applications. This lexicon not only preserves a record of Palestinian linguistic and cultural heritage but also offers a framework for computational analysis of post-Nakba literature.

Through this project, we highlight the significant yet often overlooked contributions of female voices in Palestinian literature, emphasising the importance of diverse perspectives in documenting historical trauma and resilience. The Nakba Lexicon stands as a resource for analysing the linguistic, emotional, and cultural dimensions of Palestinian literary works, enabling a more nuanced understanding of the lasting impact of the Nakba on Palestinian identity and memory.

2 Related Work

Research on the linguistic and cultural preservation within Palestinian literature has gained traction in recent years, particularly as scholars explore the ways in which language and literature document and sustain collective memory. Palestinian authors and poets have frequently employed their work as a form of resistance and preservation, capturing the personal and collective traumas associated with displacement and cultural erasure (McDonald, 2013). Much of this research underscores the role of poetry and narrative in retaining pre-Nakba identities, toponyms, and cultural references, emphasising literature as a safeguard against the loss of heritage (Uebel, 2014; El-Ghadban and Strohm, 1900).

The influence of Mahmoud Darwish is central to studies on Palestinian poetry, with his works providing foundational insights into themes of exile, memory, and identity. Darwish's poetry is widely cited as a significant source of inspiration for Palestinian authors, including Istiqlal Eid, who not only shares Darwish's geographic roots but also his commitment to documenting Palestinian heritage (Mattawa, 2014). Scholars have examined Darwish's impact on modern Palestinian literature and how his style of prose poetry has been adapted by subsequent generations of Palestinian poets who seek to articulate their own experiences of statelessness and longing (Eid, 2016).

In the field of Natural Language Processing (NLP), efforts to develop language models and datasets for low-resource languages, particularly dialects, are also relevant to this study (Magueresse et al., 2020; El-Haj et al., 2015). Research on Arabic dialects and low-resource NLP applications provides insights into the complexities of handling linguistic diversity within Palestinian literature, especially given the mixture of Modern Standard Arabic and Palestinian dialect in Eid's work (Kwaik et al., 2018; Darwish et al., 2021). Projects focusing on the creation of lexicons and domain-specific terminologies have demonstrated the potential for computational approaches to capture unique linguistic and cultural expressions, facilitating further analysis and preservation of marginalised narratives (Sonn et al., 2013).

Work on thematic and emotional lexicons has also been explored in NLP, particularly in the context of trauma and cultural resilience (Kirmayer et al., 2009). Studies have shown that by structuring lexicons around themes such as political terminology, sensory language, and cultural references, researchers can gain a more profound understanding of how language embodies historical events and communal identity (Kenter et al., 2012; Schmidt and Burghardt, 2018). The categorisation approach proposed in this study builds on these foundations by organising terms specific to the Nakba within distinct domains, enabling a targeted NLP analysis that respects the cultural context of the lexicon.

This work contributes to these existing efforts by developing a Nakba-focused lexicon within Palestinian literature, which aims to support both linguistic preservation and nuanced computational analysis. Through a female poet's perspective, our research not only enhances existing NLP applications but also addresses the gendered aspects of cultural documentation, highlighting the significant yet often underrepresented contributions of women in post-Nakba literary production.

3 Overview of the Author

Istiqlal Eid (?-), a poet originally from El-Birweh, incorporates the term "Biladna" (meaning "Our Country") into her identity, referring to herself as "Bint el-Birweh" (the Daughter of El-Birweh). Currently residing in Tamra and working as an English teacher, her name, meaning "independence," was chosen by her father in hope of Palestinian autonomy after the Nakba of 1948. Later, Eid added "Biladna" to signify her desire for freedom from oppression and corruption in Arab lands. El-Birweh, her birthplace, is also significant as it is the hometown of celebrated poet Mahmoud Darwish, her maternal uncle.

Eid identifies as a refugee in her own homeland. Her family, despite remaining in what is now Israel's 1948 territories, cannot reside in El-Birweh, which has since become a Jewish settlement. Eid's main publication, a poetry collection titled *ʾĀsifa ᶜAlā el-ʾIzᶜāj* (آسِفة على الإزعاج, *Sorry for the Disturbance*), was released in 2017. She is also working on a collection of short stories titled *Šḥār wimšaḥāra* (شحار و مشحَّرة, *Smears and a Smeared Woman*).

Our research analyses selections from Eid's 2017 *Diwan*, rooted in the cultural and geographical exile reminiscent of Darwish. Her family's displacement to Tamra left them labelled as "present absentees[3]" (Makhoul, 2012), indicating their presence in the state yet denial of access to ancestral lands.

Eid's prose poetry style often adopts a sarcastic tone, written in a blend of Modern Standard Arabic and Palestinian dialect. Her work reflects an effort to safeguard the Palestinian narrative, documenting pre-Nakba names, figures, and places, and preserving Palestinian history and Nakba memories.

Arabic Language

The Arab world has long experienced a state of *diglossia*, defined as the simultaneous presence of two distinct levels of language, or even two different languages, within the same society, each occupying distinct communicative domains (Ferguson, 1959; Fishman, 1967). In the Arabic-speaking world, this results in the coexistence of vernacular Arabic, or *dialectal Arabic*—a collection of localized, often mutually unintelligible varieties—and *Modern Standard Arabic* (MSA), the standardized, literary form used in formal settings such as media, education, and interregional communication. MSA is taught in schools and represents the unifying linguistic thread across Arabic-speaking regions, though it remains separate from daily spoken varieties (Owens, 2001; Versteegh, 1997).

These dialectal forms of Arabic vary significantly across regions, with a sort of linguistic continuum existing along geographical lines. Generally, dialects become more divergent as geographical distance increases; for instance, the Arabic spoken in Galilee closely resembles the dialect of southern Lebanon, while eastern Moroccan Arabic is more akin to western Algerian varieties. Yet dialect similarity also depends on historical, social, and religious factors, which create distinct links between urban and rural varieties. The Arabic spoken in Jerusalem, for instance, shares specific features with that of Beirut, as both are urban dialects with complex social histories (Rosenhouse, 2007). Additionally, many dialects in the Arab world are influenced by contact with other languages, which further diversifies the linguistic landscape. In Iraq, Arabic dialects have absorbed elements from Aramaic, Kurdish, and Farsi, while in North Africa, Tamazight influences are prominent (Al-Wer and de Jong, 2009).

4 The Linguistic Context of Palestinians in Israel

The declaration of Israel in 1948 reshaped not only the territorial but also the linguistic landscape, transforming Palestinian citizens from a majority into a minority with a marginalized language. The shift placed Arabic in a subordinate position relative to Hebrew, reflecting the asymmetrical political and social power dynamics between the Palestinian minority and the Jewish Israeli majority (Henkin-Roitfarb, 2011). After 1948, Arabic was increasingly perceived within Israel as a language of opposition,

[3]According to Makhoul (2012) in their Survey of Palestinian Refugees and Internally Displaced Persons 2013-2015 (p. 8), internally displaced persons (IDPs) in Mandate Palestine fall into two main categories. The first group includes approximately 384,200 Palestinians who have been displaced within Israel since 1948, while the second group comprises around 334,600 Palestinians displaced within the territories occupied since 1967

while Hebrew was elevated as the cornerstone of the nation-building process (Spolsky and Shohamy, 1999).

For Palestinian citizens of Israel, proficiency in Hebrew became essential for navigating official and social settings, while Arabic faced reduced support in public institutions. In Jewish Israeli schools, Arabic instruction was often limited to military contexts, highlighting the asymmetrical status of the two languages (Spolsky and Shohamy, 1999; Amara, 2002). This bilingual reality reflects a linguistic hierarchy, where Arabic serves as both a practical language and a symbol of cultural resilience. In literature and cultural expression, language plays a critical role in maintaining and asserting Palestinian identity. Despite its marginalized status, Arabic serves as a medium for exploring themes of identity, resistance, and cultural continuity.

Given these influences, the Arabic of the '48 Palestinians occupies a distinct position in the landscape of Palestinian dialectology. The continuous interaction with Hebrew and the isolation from other Palestinian dialects contribute to a rich, hybrid linguistic identity that reflects the historical and social complexities of Palestinian communities within Israel (Horesh, 2021).

5 Methodology and Dataset Classification

Eid's work provides a rich source of Nakba-related terminology, facilitating the creation of a comprehensive dataset for Natural Language Processing (NLP) applications. By focusing on a female voice, we aim to create an inclusive dataset that captures diverse experiences in Palestinian literature, while emphasising the significant yet often overlooked contributions of women. This approach enables nuanced NLP analyses of language and themes in post-Nakba literary works.

The dataset is structured into thematic categories, capturing linguistic adaptations to evolving Palestinian identity and socio-political realities:

1. Political Terminology

Terms in this category describe new political realities post-Nakba, such as "peace,", "war," "occupation," and "resistance," as well as names of displaced communities, refugee camps, and geopolitical terms resulting from the conflict.

2. Memory and Preservation

This category encompasses literary efforts to preserve the pre-Nakba past, highlighting key events, significant dates, and collective tragedies. It includes terms commemorating losses, displacement, and destruction of Palestinian communities, documenting shared trauma and the struggle for remembrance of pre-1948 Palestine.

3. Sensory and Emotional Lexicon

Words conveying sensory experiences and emotions such as pain, loss, displacement, and longing, evoke the physical and psychological impact of the Nakba on individuals and communities.

4. Toponyms and Place Names

This category records Palestinian names of cities, villages, and regions from both pre-1948 and post-Nakba periods. It reflects the geographical and cultural evolution caused by conflict and occupation.

5. Names of People

Palestinian activists resist the reduction of their identities to numbers—whether as war casualties or UNRWA[4] ration card recipients. Palestinian authors counteract this by recording names of those who died, thus our dataset compiles names from Eid's *Diwan*, both of well-known figures and ordinary people.

6. Social and Cultural Lexicon

Terms here relate to traditional customs, songs, proverbs, and cultural practices that embody Palestinian identity. The lexicon reflects the continuity and transformation of cultural expressions from pre- to post-Nakba.

[4]UNRWA stands for the United Nations Relief and Works Agency for Palestine Refugees in the Near East. It was established in 1949 to provide assistance and protection to Palestinian refugees displaced during the 1948 Arab-Israeli conflict and subsequent wars. UNRWA offers services such as education, healthcare, social services, and emergency aid in its areas of operation, which include the West Bank, Gaza Strip, Jordan, Lebanon, and Syria.

7. Natural World

This category includes names of plants, trees, herbs, and animals significant to the Palestinian landscape, often symbolising resilience and rootedness. Some plants were uprooted post-Nakba, with names changing due to dialectal influences in Lebanon, Syria, and Jordan. Eid preserves these terms in the Palestinian memory archive.

8. Hebrew and Linguistic Influence

Addressing the forced incorporation of Hebrew terminology in Palestinian literature due to occupation, this category includes terms borrowed from Hebrew or used to describe life under occupation.

This classification structure illustrates the diversity of linguistic responses to historical, cultural, and political shifts surrounding the Nakba. It underscores the efforts to preserve cultural memory, adapt to new realities, and articulate experiences of displacement and resistance. For specific examples of expressions and their classifications, see Appendix A.

6 Experiments

Building upon the methodology and thematic classification discussed earlier, we conducted experiments to analyse the linguistic and cultural nuances embedded in the dataset. The dataset comprises 222 sentences, each carefully annotated with a word or phrase that reflects its language type and lexical class. These annotations aim to capture the complex interplay of linguistic elements present in Istiqlal Eid's work, including Modern Standard Arabic, Palestinian dialect, and Hebrew influences.

The annotated data is categorised into ten distinct lexical classes, as illustrated in Figure 1. The distribution highlights the diversity of terms used in post-Nakba literature, with notable proportions allocated to categories such as Toponyms (27.5%), Names of People (21.7%), and Politics (18.3%). Smaller but significant categories include Culture (9.2%), Nature (7.9%), Memory (7.9%), and Sensory and Emotional Lexicon (5.0%). External linguistic influences, including Hebrew (1.3%), French (0.8%), and English (0.8%), reflect the historical and sociopolitical interactions shaping the linguistic landscape.

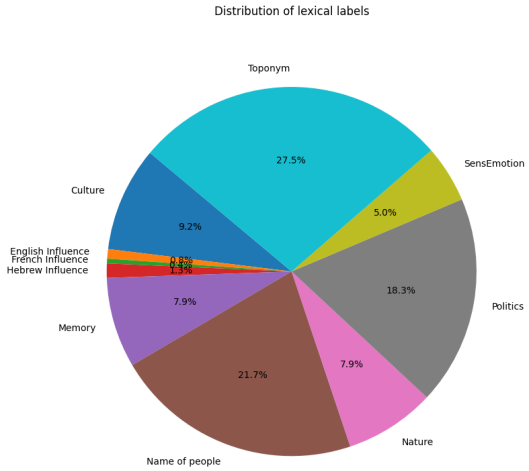

Figure 1: Lexical labels

These classifications provide insight into the thematic emphasis of the Nakba Lexicon, illustrating how language is used to articulate identity, resistance, and memory. To complement this analysis, each sentence has also been identified in terms of its language type, classified into six categories, as detailed in Table 1. This dual classification highlights the dynamic interaction of standardised and dialectal forms, alongside the incorporation of external linguistic influences, offering a holistic view of the dataset's linguistic diversity.

This experiment provides a foundation for exploring how language is used in Palestinian literature to articulate themes of displacement, memory, and resilience, while also enabling computational analysis of these themes within Natural Language Processing applications.

Label	Count	Percentage
Dialect	35	15.77%
Hebrew (translitterated)	2	0.90%
Neologism - Standard	1	0.45%
Standard	185	83.33%
Standard DialectStandard	1	0.45%
Standardized neologism(means being Israelised)	1	0.45%

Table 1: Language type distribution

7 Lexical Classification

To further investigate the linguistic richness and thematic organisation of the dataset, we applied machine learning techniques for lexical classification. The goal of this step was to assess how effectively computational models can identify and categorise the nuanced lin-

guistic elements present in Istiqlal Eid's work, particularly across Modern Standard Arabic, Palestinian dialect, and Hebrew influences.

The primary technique employed for this task was the Term Frequency-Inverse Document Frequency (TF-IDF) embedding approach, which represents texts as numerical vectors based on their significance within the dataset. These vectors were used as input to train four traditional machine learning models: Support Vector Machines (SVM), Logistic Regression, Random Forest, and Naïve Bayes. These models were selected for their proven effectiveness in text classification tasks, where SVM excels at finding optimal hyperplanes for classification, Logistic Regression provides interpretability, Random Forest is robust against overfitting, and Naïve Bayes is efficient for small datasets. To address the limited size of the dataset, bigram features were incorporated (`ngram_range=(1, 2)`) to enhance representation by capturing contextual relationships between words.

In addition to these traditional approaches, two state-of-the-art pre-trained language models, ARABERT (Antoun et al., 2020) and MARBERT (Abdul-Mageed et al., 2021), were utilised to classify the lexical classes. These models were chosen for their specialised design, which caters to the unique linguistic characteristics of Arabic, including morphology, syntax, and dialectal variations. ARABERT is tailored for Modern Standard Arabic tasks, while MARBERT focuses on Arabic dialects, making both models well-suited to handle the mix of Standard Arabic, Palestinian dialect, and Hebrew influences present in the dataset. Both models were evaluated in two settings: first, as-is without additional training, to assess their generalisation capabilities, and second, fine-tuned on our dataset to adapt them to the specific nuances of post-Nakba literature. Fine-tuning was conducted with early stopping criteria to prevent overfitting, halting training after three consecutive evaluation epochs without validation loss improvement. Model checkpoints were saved after each epoch to ensure the best possible performance.

Preprocessing was applied to standardise the input data, including steps such as diacritisation removal, extra space trimming, stop word elimination, symbol cleaning, and character normalisation. These steps ensured the dataset was consistent and suitable for effective computational analysis.

This classification effort not only provides insights into the dataset's linguistic diversity but also highlights the potential of machine learning and large language models in analysing post-Nakba literature. By bridging traditional and modern NLP techniques, the experiments showcase a comprehensive approach to understanding and preserving Palestinian linguistic and cultural heritage.

8 Results and Discussion

The experimental results highlight the linguistic diversity captured in the Nakba Lexicon dataset, which is essential for understanding the interplay of language types in post-Nakba literature. Table 1 provides a breakdown of the language types annotated within the dataset. Standard Arabic constitutes the majority (83.33%), reflecting its role as the primary medium for formal literary expression. Dialectal Arabic, accounting for 15.77% of the dataset, highlights the significance of regional vernaculars in conveying personal and cultural narratives. Smaller contributions, such as transliterated Hebrew (0.90%), illustrate the linguistic influence of socio-political contexts, particularly the interaction between Palestinian Arabic and Hebrew. This distribution demonstrates the dataset's potential for analysing both the standardised and dialectal aspects of Arabic, as well as cross-linguistic interactions in Palestinian literature.

The lexical classification task assesses the ability of models to categorise words and phrases into predefined lexical classes, reflecting the diverse linguistic elements of the Nakba Lexicon. Table 2 summarises the performance of various machine learning models on this task. Traditional models, such as SVM and Random Forest, yielded moderate results, with F1-scores ranging from 0.09 to 0.19, indicating limitations in capturing the complexity of the dataset. In contrast, large language models (LLMs) such as ARABERT and MARBERT demonstrated significant improvements. Fine-tuned MARBERT achieved the highest F1-score of 0.68, showcasing its ability to effectively capture and classify the nuanced lexical

features present in Arabic literature. These results highlight the importance of domain-specific pre-trained models for processing culturally and linguistically rich datasets like the Nakba Lexicon.

The language classification task evaluates the ability of models to identify the language type of each sentence within the dataset, reflecting its multilingual and dialectal diversity. Table 3 presents the performance of various models in this task. Traditional machine learning models, including SVM, Logistic Regression, and Random Forest, provided consistent and reasonable results, achieving F1-scores in the range of 0.76 to 0.77. These models, however, lacked the sophistication needed to fully capture the complexity of multi-class classification in Arabic datasets. Fine-tuned ARABERT outperformed all other models, achieving the highest F1-score of 0.87, demonstrating its capability to handle intricate linguistic variations and multi-class tasks effectively. MARBERT closely followed with an F1-score of 0.84, further validating the efficacy of pre-trained language models in addressing the challenges posed by diverse linguistic datasets like the Nakba Lexicon.

Models	**Precision**	**Recall**	**F1-score**
SVM	0.36	0.13	0.19
Logistic Regression	0.27	0.05	0.09
Random Forest	0.34	0.11	0.17
Naive Bayes	0.41	0.11	0.17
AraBERT (10 epoch)	0.70	0.42	0.53
MARBERT (7 epoch)	**0.88**	**0.56**	**0.68**

Table 2: Performance of Machine Learning and Pre-trained Models in Lexical Classification Tasks

These results underscore the profound challenges posed by the linguistic diversity and nuanced language use in post-Nakba literature. The Nakba Lexicon, as reflected in the dataset's composition and classification tasks, captures a unique confluence of standardised, dialectal, and externally influenced linguistic features. This complexity mirrors the fragmented identities and cultural resilience of Palestinians, as expressed in their literary and linguistic heritage. The moderate performance of traditional machine learning models highlights the difficulty of computationally analysing such a linguistically rich and context-dependent dataset. These models, while capable of providing baseline insights, struggle to fully grasp the intricate layers of meaning and cultural references embedded in post-Nakba narratives.

Conversely, the superior performance of fine-tuned large language models, such as ARABERT and MARBERT, illustrates their capacity to bridge computational methods with cultural and historical contexts. By leveraging domain-specific pre-trained models, this study demonstrates how advanced NLP techniques can contribute to preserving and analysing Palestinian narratives in a way that respects their linguistic and cultural particularities. The high F1-scores achieved by ARABERT and MARBERT, particularly in language classification, underscore their potential to capture the interplay of Modern Standard Arabic, Palestinian dialects, and Hebrew influences within Nakba-related literature. This capability is crucial for understanding how language has been used as a medium of resistance, memory preservation, and identity formation in the face of displacement and marginalisation.

Furthermore, these findings highlight the value of computational approaches in elevating underrepresented narratives in global discourses. By enabling the analysis of low-resource languages and culturally rich datasets, NLP offers a pathway to amplify marginalised voices and ensure their stories are preserved for future generations. In the case of the Nakba Lexicon, this study not only contributes to the documentation of Palestinian heritage but also lays the groundwork for applying similar methods to other marginalised linguistic communities.

Models	**Precision**	**Recall**	**F1-score**
SVM	0.71	0.82	0.76
Logistic Regression	0.71	0.84	0.77
Random Forest	0.71	0.84	0.77
Naive Bayes	0.71	0.84	0.77
AraBERT (7 epoch)	**0.87**	**0.87**	**0.87**
MARBERT (8 epoch)	0.84	0.84	0.84

Table 3: Performance of Machine Learning and Pre-trained Models in Language Classification Tasks

Ultimately, this research illustrates the potential of integrating computational tools with literary and cultural studies to address historical traumas and support cultural resilience. The challenges faced in classifying such a nuanced dataset reflect broader issues in preserving endangered narratives, while the successes

achieved with advanced models point to a future where technology can play a vital role in safeguarding cultural memory. The Nakba Lexicon serves as both a technical achievement and a testament to the enduring power of language in articulating collective experiences of resistance, loss, and hope.

9 Conclusion and Future Work

This study introduced the Nakba Lexicon, a comprehensive dataset derived from the poetic works of Istiqlal Eid, capturing the linguistic, cultural, and emotional dimensions of Palestinian post-Nakba literature. By categorising the dataset into thematic domains and leveraging computational methods, we demonstrated how advanced Natural Language Processing (NLP) techniques can preserve and analyse marginalised linguistic and cultural narratives. The lexicon serves as a bridge between computational tools and cultural studies, offering a resource for exploring the nuanced interplay of identity, memory, and resilience in the face of historical trauma.

The experimental results underline the challenges inherent in processing such a linguistically rich and contextually complex dataset. Traditional machine learning models, while providing baseline insights, struggled to fully capture the intricate dynamics of Palestinian literature, especially the interplay of Modern Standard Arabic, Palestinian dialects, and Hebrew influences. In contrast, pre-trained language models like ARABERT and MARBERT significantly outperformed these models, with MARBERT excelling in lexical classification and ARABERT achieving state-of-the-art results in language classification. These findings highlight the potential of domain-specific models to address the unique demands of datasets that blend linguistic, cultural, and historical layers.

The Nakba Lexicon is more than a computational dataset; it is a testament to the enduring power of language as a medium for resistance, memory preservation, and cultural continuity. By amplifying the voices embedded in Palestinian literature, this research not only contributes to the documentation of Palestinian heritage but also underscores the role of NLP in safeguarding endangered narratives. These efforts align with a broader vision of using technology to amplify the stories of marginalised communities and preserve their cultural identities for future generations.

Future work will expand the Nakba Lexicon to include additional texts from Palestinian authors and poets, aiming to enhance its representativeness and robustness. We also intend to explore cross-dialectal adaptations, capturing the linguistic diversity across Arabic-speaking regions affected by the Nakba. Integrating context-aware models, such as transformer-based architectures, will enable deeper analysis of the interplay between language, culture, and history. Furthermore, collaborative efforts with historians, linguists, and cultural preservationists will enrich the lexicon's applications, fostering interdisciplinary approaches to understanding and preserving cultural heritage.

Once this paper is accepted, the Nakba Lexicon will be released publicly, providing researchers and practitioners with a valuable resource for computational analysis and a meaningful contribution to preserving the narratives of resilience and resistance within Palestinian literature.

10 Limitations

While this study highlights the potential of the Nakba Lexicon for preserving and analysing linguistic and cultural narratives, several limitations must be acknowledged. First, the dataset size is relatively small, consisting of 222 annotated sentences, which limits the generalisability of the experimental results. While the inclusion of fine-tuned models like ARABERT and MARBERT significantly improved performance, a larger and more diverse dataset is required to ensure broader applicability and robust generalisation across different contexts and linguistic variations.

Second, the dataset primarily focuses on the works of a single poet, Istiqlal Eid, which, while rich and representative of certain aspects of Palestinian literature, may not fully capture the breadth of linguistic and cultural diversity present in the wider corpus of Palestinian writing. Expanding the dataset to include other authors, genres, and dialects would provide a more comprehensive representation of the post-Nakba narrative.

Third, the reliance on pre-trained models like ARABERT and MARBERT, while beneficial, highlights challenges in adapting NLP tools to datasets that mix standardised and dialectal Arabic, as well as incorporating influences from Hebrew and other languages. Future work should address these challenges by developing more targeted models that can better accommodate such linguistic complexities.

Lastly, the dataset's cultural and historical sensitivity requires careful handling to ensure its appropriate use in research and applications. Collaborative efforts with cultural preservationists and community stakeholders are essential to ensure that the dataset is used responsibly and meaningfully.

Despite these limitations, the Nakba Lexicon represents a significant step forward in combining computational tools with cultural preservation, offering valuable insights into the intersections of language, identity, and historical trauma. Addressing these limitations in future work will further enhance its value as a resource for interdisciplinary research.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Enam Al-Wer and Rudolf de Jong. 2009. *Arabic Dialectology: In Honour of Clive Holes on the Occasion of his Sixtieth Birthday*. BRILL.

Muhammad Amara. 2002. The place of arabic in israel.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R El-Beltagy, Wassim El-Hajj, et al. 2021. A panoramic survey of natural language processing in the arab world. *Communications of the ACM*, 64(4):72–81.

Istiqlal Eid. 2017. *Asifa 'Ala al-Iz'aj [Sorry for the Disturbance]*. Dar al-Arkan lil-Intaj wa al-Nashr, Israel.

Muna Abu Eid. 2016. *Mahmoud Darwish: literature and the politics of Palestinian identity*. Bloomsbury Publishing.

Yara El-Ghadban and Kiven Strohm. 1900. The ghosts of resistance: dispatches from palestinian art and music. *Palestinian music and song expression and resistance since*, pages 175–200.

Mahmoud El-Haj, Udo Kruschwitz, and Chris Fox. 2015. Creating language resources for under-resourced languages: methodologies, and experiments with arabic. *Language Resources and Evaluation*, 49:549–580.

Charles A. Ferguson. 1959. Diglossia. *Word*, 15(2):325–340.

Joshua A. Fishman. 1967. Bilingualism with and without diglossia; diglossia with and without bilingualism. *Journal of Social Issues*, 23(2):29–38.

Roni Henkin-Roitfarb. 2011. Hebrew and arabic in asymmetric contact in israel. *Lodz Papers in Pragmatics*, 7(1):61–100.

Uri Horesh. 2021. Palestinian dialects and identities shifting across physical and virtual borders. *Multilingua: Journal of Cross-Cultural and Interlanguage Communication*, 40(5):647–673.

Tom Kenter, Tomaž Erjavec, Maja Žorga Dulmin, and Darja Fišer. 2012. Lexicon construction and corpus annotation of historical language with the CoBaLT editor. In *Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 1–6, Avignon, France. Association for Computational Linguistics.

Laurence J Kirmayer, Megha Sehdev, Rob Whitley, Stéphane F Dandeneau, and Colette Isaac. 2009. Community resilience: Models, metaphors and measures. *International Journal of Indigenous Health*, 5(1):62–117.

Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. 2018. Shami: A corpus of levantine arabic dialects. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

Manar H Makhoul. 2012. The forces of presence and absence: Aspects of palestinian identity transformation in israel between 1967-1987. *Journal of Levantine Studies*, 2(1):135–159.

Khaled Mattawa. 2014. *Mahmoud Darwish: The poet's art and his nation*. Syracuse University Press.

David A McDonald. 2013. *My voice is my weapon: Music, nationalism and the poetics of palestinian resistance*. Duke University Press.

Jonathan Owens. 2001. *Arabic as a Minority Language*. Walter de Gruyter.

Judith Rosenhouse. 2007. The arabic dialects in the north of israel. *Bulletin of the School of Oriental and African Studies, University of London*, 55(1):19–41.

Ahmad H Sa'di and Lila Abu-Lughod. 2007. *Nakba: Palestine, 1948, and the claims of memory*. Columbia University Press.

Thomas Schmidt and Manuel Burghardt. 2018. An evaluation of lexicon-based sentiment analysis techniques for the plays of Gotthold Ephraim Lessing. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 139–149, Santa Fe, New Mexico. Association for Computational Linguistics.

Christopher C Sonn, Garth Stevens, and Norman Duncan. 2013. Decolonisation, critical methodologies and why stories matter. In *Race, memory and the apartheid archive: Towards a transformative psychosocial praxis*, pages 295–314. Springer.

Bernard Spolsky and Elana Shohamy. 1999. *The Languages of Israel: Policy, Ideology, and Practice*. Multilingual Matters.

Carly M Uebel. 2014. *Carrying Palestine: Preserving the" Postmemory" Palestinian Identity and Consolidating Collective Experience in Contemporary Poetic Narratives*. Ph.D. thesis, University of Oregon.

Kees Versteegh. 1997. *The Arabic Language*. Edinburgh University Press.

A Appendix: Expressions and Lexical Examples

This appendix provides detailed examples of expressions and lexical terms from the Nakba Lexicon, including linguistic and cultural explanations.

1. **16 / 1 / Dialect / Sensory Name of Place:** أحمرالفاير
   aḥmar al-fayir "blazing red" A dialectal expression composed of أحمر aḥmar 'red' and the participle فاير fayir 'burning, raging'. Fayir is a local dialectal pronunciation related to Standard Arabic (SA) فائر fā ir. The Palestinian Arabic (PA) has dropped the hamza and replaced it by the consonant /y/, aligning it with the vowel /i/. Comparison: SA فائر fā ir - PA فاير fayir.

2. **19 / 1 / Dialect / Hebrew Influence:** السيكوباتية
   Sikopatiya "Psychopathy" The form *sikopatiya* may derive from SA سيكوباتية sīkūbātiyya / saykūbātiyya or from English "psychopathy," with phonological influence from Modern Hebrew (MH). The Palestinian Jordanian variant seems to retain English features such as the diphthong /ay/ and interdental /t̲/.

3. **21 / 2 / Dialect / Cultural (Dialect):** إحنا بخير
   iḥna bxer "We're all right" Common PA form إحنا بخير iḥna b-xer compared with SA نحن بخير naḥnu bi-xayr. PA features include the dropping of short vowels (بـ b-, cp. SA بـ bi-), diphthong reduction (خير xer, SA خير xayr), and a specific form for the first-person personal pronoun إحنا iḥna (SA نحن naḥnu).

4. **28 / 2 / Dialect / Hebrew Influence:** هلّولويا هلّولويا
   Halluluya halluluya "Hallelujah, hallelujah" Originates from Hebrew *halelu yah* "praise God," although it is difficult to determine when it was borrowed. Possibly through Christianity, making it a very ancient legacy, but it might have been updated through contact with MH.

5. **32 / 1 / Dialect / Hebrew Influence:** فنطازيا
   Fanṭazya "fantasy" A local adaptation of the international word "fantasy." Unlike usual SA فانتازيا fāntāzya, the PA form includes an emphatic /ṭ/, i.e., *fanṭazya*, which might be an internal evolution or, less probably, an orthographic influence from MH.

6. **36 / 5 / Dialect / Cultural: Proverb:** على أد فراشك مدّ اجريك
ala ədd frašak mid ižrik "Cut your coat according to your cloth." This PA proverb contrasts with SA: PA على أد ala ədd vs. SA على قد alā qadd. Additionally, metathesis is seen: PA اجر ižr vs. SA رجل rižl.

7. **140 / 2 / Dialect / Toponym:** صباح الخير على مخيّم شاتيلا
Ṣabaḥ al-xer ala muxayyam šātīlā "Good morning Shatila refugee camp" The word شاتيلا Šātīlā might be related to the Aramaic root *√š.t.l.* related to "plants," as seen in the Syriac words *štilta* 'plantation' and *štala* 'to plant.'

8. **130 / 1 / Standardised Neologism / Political Register (and Cultural): Post-Nakba:** أسرلتنا
Asralatna A verb indirectly derived from the name 'Israel,' from which the consonant root *√s.r.l.* has been extracted. This neologism demonstrates the ability of Semitic linguistic structures to create roots from foreign words.