Arabic Topic Classification Corpus of the Nakba Short Stories

Osama Hamed[1,*] and Nadeem Zaidkilani[2,3,*]

[1]Department of Computer Systems Engineering, Palestine Technical University - Kadoorie, Palestine
osama.hamed@ptuk.edu.ps, sam.hamed@gmail.com
[2]Department of Computer Engineering and Security Mathematics, University Rovira i Virgili, Spain
[3] Department of Engineering and Technology, Al-Zaytona University of Science and Technology, Palestine
nadeem.zaidkilani@estudiants.urv.cat, nadeem.kilani@zust.edu.ps,nadimkilani@gmail.com

[*]These authors contributed equally to this work.

Abstract

In this paper, we enrich Arabic Natural Language Processing (NLP) resources by introducing the "Nakba Topic Classification Corpus (NTCC)," a novel annotated Arabic corpus derived from narratives about the Nakba. The NTCC comprises approximately 470 sentences extracted from eight short stories and captures the thematic depth of the Nakba narratives, providing insights into both historical and personal dimensions. The corpus was annotated in a two-step process. One third of the dataset was manually annotated, achieving an IAA of 87% (later resolved to 100%), while the rest was annotated using a rule-based system based on thematic patterns. This approach ensures consistency and reproducibility, enhancing the corpus's reliability for NLP research. The NTCC contributes to the preservation of the Palestinian cultural heritage while addressing key challenges in Arabic NLP, such as data scarcity and linguistic complexity. By like topic modeling and classification tasks, the NTCC offers a valuable resource for advancing Arabic NLP research and fostering a deeper understanding of the Nakba narratives

1 Introduction

Automatic document categorization has gained significant importance due to the continuous influx of textual documents on the web. The rise of the Internet and Web 2.0 has led to an unprecedented increase in unstructured data generated from various sources, particularly social media. This vast array of unstructured information presents both a challenge and an opportunity for data processing and management, enabling researchers to extract meaningful insights. One of the key tasks in this realm is text classification, which has witnessed substantial advancements recently, particularly with the advent of machine learning (ML) techniques (Elnagar et al., 2020).

Text categorization, often referred to interchangeably as text classification, involves predicting predefined categories or domains for a given document. This automated process can either identify the most relevant single category or multiple closely related categories. Given the enormous volume of available documents online, manual classification is impractical, necessitating automated classifiers that transform unstructured text into machine-readable formats (Elnagar et al., 2020).

While text categorization has been well-studied in several languages, including English, the Arabic language remains underrepresented in this research area. Despite Arabic being the fourth most widely used language on the Internet and the sixth official language recognized by the United Nations (Wahdan et al., 2024), there are few studies focusing on Arabic text classification (Alyafeai et al., 2022). The scarcity of comprehensive and accessible Arabic corpora presents significant obstacles for researchers. Most existing datasets are small, lack predefined classes, or require extensive modifications before use. This limitation complicates the validation and comparison of proposed methods, hindering progress in Arabic text categorization (Elnagar et al., 2020).

In this paper, we introduce the NTCC, a new Arabic dataset specifically designed for emotion detection in narratives surrounding the Nakba—a pivotal event in Palestinian history characterized by displacement and loss. These stories are crucial for preserving the historical events and documenting the suffering of the Palestinian people since 1948. Our dataset encompasses five distinct categories: (i) historical events and politics, (ii) emotions and spirituality, (iii) nature and daily life, (iv) homesickness and war/conflict, and (v) others. "Others" is dedicated to sentences that do not belong to any of the defined categories. This structure aims to provide researchers with flexibility in annotation and a more nuanced understanding of the narratives.

Proceedings of the Proceedings of the 1st International Workshop on Nakba Narratives as Language Resources, pages 48–55
January 20, 2025. ©2025 Association for Computational Linguistics

By releasing this dataset, which consists of approximately 470 sentences extracted from eight stories, we aim to facilitate the application of various NLP and machine learning related tasks. Our Arabic annotated corpus will pave the way for researchers in machine learning and NLP to conduct numerous studies, potentially leading to advancements in sentiment analysis, topic modeling, and other classification tasks. Notable works in this area include the use of deep learning models for text classification, sentiment analysis using recurrent neural networks (RNNs), and transformer-based architectures such as BERT for enhanced context understanding.

This work contributes to preserving Palestinian heritage and the Palestinian issue by documenting these experiences. Our objective is to enhance the predictive capabilities for semantic analysis on future unseen data, while contributing to the growing body of research in Arabic corpora and text classification.

This paper is organized as follows. In Section 2, we present prior and recent research on Arabic topic classification. Section 3 provides a comprehensive analysis of the data handling. Section 4 summarizes the steps followed, and describes the proposed approach for corpus construction. In Section 5, we conclude and point out ideas for future search.

2 Related Work

The construction and utilization of structured Arabic corpora are essential for advancing Arabic text classification and Natural Language Processing (NLP). Due to the unique challenges posed by Arabic—including its complex morphology, rich dialectal variations, and limited open-source resources—the development of specialized corpora has become a focal point in recent studies.

Albared et al. (2023) propose an approach to Arabic topic classification using generative and AutoML techniques, demonstrating how the success of classification models is heavily dependent on high-quality, diverse datasets. This study underscores the necessity of large, labeled corpora that allow models to generalize effectively, addressing Arabic's unique linguistic challenges. In a similar direction, the OSAC (Open-Source Arabic Corpora) initiative (Saad and Ashour, 2010) tackles resource scarcity by providing open-access corpora to improve classification and clustering. OSAC compiles a wide range of Arabic texts, including social media and news articles, thereby supporting the development of more robust classification models.

In line with OSAC, the Ahmed and Ahmed (2021) study on Arabic news classification develops a specialized corpus that enhances machine learning algorithms for news categorization. It stresses the importance of balanced data across categories to mitigate classification biases. Likewise, systematic reviews such as (Elnagar et al., 2020) identify trends across Arabic NLP resources, noting that existing corpora often suffer from domain specificity and dialectal homogeneity. These studies collectively advocate for more diversified corpora that cover both Modern Standard Arabic (MSA) and regional dialects to bolster model robustness across different language forms.

Addressing specific NLP challenges, AL-Sarayreh et al. (2023) discuss data augmentation and annotation techniques as solutions to enrich Arabic corpora, especially in low-resource contexts. They suggest that creative approaches to annotation and corpus expansion are essential for capturing Arabic's linguistic diversity. Furthermore, Wahdan et al. (2024) emphasizes the importance of domain-specific corpora and advocates for expanding corpus types to represent the broader spectrum of Arabic content. This work argues that larger and more varied datasets can yield substantial improvements in classification accuracy and task transferability.

In summary, as Arabic NLP research progresses, the development of specialized, open-source, and diverse Arabic corpora remains critical. The aforementioned studies contribute to this goal by offering resources that not only enhance classification accuracy, but also address broader challenges related to language variation, resource availability, and domain-specific needs in Arabic NLP.

3 Data Handling

We want to construct a new Arabic topic classification corpus based on the Nakba narratives. The raw data represents a set of Nakba short stories that contains the narratives of Nakba, e.g. the suffering, memories, ... etc. as they were told by refugees. The stories were taken from the Nakba Archive website[1].

[1] https://www.nakba-archive.org

3.1 Raw Data Gathering

On the Nakba Archive website, in the project section, we have extracted a collection of eight Nakba stories, written by different authors, which are in PDF format. Each of these stories needs a special attention to become suitable for NLP related tasks.

3.2 Data Cleaning

In addition to the Arabic diacritics that rarely appear in the text. We noticed that the stories contain some special characters, a few English texts, and web references as uniform resource locators (URLs). The diacritics were removed using the PyArabic (Zerrouki, 2023) Python library. Whereas, the remaining noise was removed using regular expressions.

3.3 Data Preprocessing

We want to create/construct an annotated dataset for topic classification that is taken from Nakba stories. We started this task by data preprocessing, which falls into three main steps:

Convert PDF to Structured Format Although the PDF is primarily a format for visual presentation, it contains unstructured data. Instead, we decided to convert to a structured format like CSV or Excel. This not only makes it easy to work with NLP tools/libraries such as Pandas and NLTK (Bird et al., 2009), but also allows any additional metadata to be stored alongside the text.

Text Normalization With the help of PyArabic, we normalized the Arabic text by removing the extra white spaces and the Tatweel.

Paragraph to Sentences Each paragraph was split into sentences based on appropriate sentence ending using the **NLTK** Python library. This will enable us to annotate the sentences easily, and add the corresponding labels in the future. The eight stories result in a set of 605 sentences. As part of the preprocessing stage, meaningless sentences—such as incomplete phrases, non-informative lines (e.g., "etc."), or formatting artifacts—were manually identified and removed by annotators. This process ensured that the dataset, which ultimately consisted of 473 contextually relevant sentences, was clean and ready for further analysis.

3.4 Topic Categorization

To categorize the topics covered in each story found on the Nakba archive website, two experts analyzed the text and listened to the interviews made with refugees. The experts were able to classify the Nakba stories' topics into four main categories. We also added the "Other" category to avoid mismatch in classification.

- Historical Events and Politics: This category includes key historical and political events of the Nakba, including key events, political decisions, and their implications on Palestinian society.

- Emotions and Spirituality: This category includes narratives that express deep emotional experiences such as grief, loss, and hope, alongside spiritual reflections.

- Nature and Daily Life: This category merges the depiction of the natural Palestinian landscape with the rhythms of everyday life.

- Homesickness and War/Conflict: This category covers the emotional longing for a lost homeland and the harsh realities of war and conflict. It combines narratives that express a deep sense of nostalgia and displacement with stories that highlight the struggles and violence experienced during the Nakba, emphasizing the enduring impact of conflict on individuals and communities.

Table 1 contains relevant examples of those five categories. We are going to use these categories as a basis for creating and annotating the Nakba Arabic topic classification corpus.

3.5 Corpus Category-Based Stats and Visualization

This work contributes to preserving Palestinian heritage by documenting these experiences, aligning with recent efforts in NLP for cultural preservation (Cabezas et al., 2022).

Nakba Stories Stats We are working with a relatively small dataset of stories. Table 2 shows the "No. of Tokens" for the eight Nakba stories.

Nakba Stories Word-Cloud The word-cloud in Figure 1 visually represents the most frequent terms

Category	Arabic Keywords Examples	Translation
Historical Event & Politics	سياسات قمعية ، النكبة	Nakba, Repressive Policies
Emotions & Spirituality	مشاعر الحزن، الإيمان	Faith, Sadness Feelings
Nature & Daily Life	شروق الشمس، العمل في الحقول	Working in the Fields, Sunrise
Homesickness & War-Conflict	حلم العودة ، قصف	Bombing, Dream of Return

Table 1: The topic categories found in Nakba stories.

ID	Arabic Title	English Title	No. of Tokens
1	افتتاحية	Introduction	556
2	ذاكرة لا تفنى	An Enduring Memory	1472
3	طوابير	Queues	2082
4	أبو عادل الفاتح	Abu Adel, The Opener	2069
5	حنين	Longing	1113
6	الدخان فالخبز فالأسلاك الشائكة	Smoke, bread, & Barbed Wire	1479
7	أم الشهيد	The Martyr's Mother	1700
8	كيف لي أن أغفر	How Can I Forgive?	1817

Table 2: The statistics of the Nakba stories.

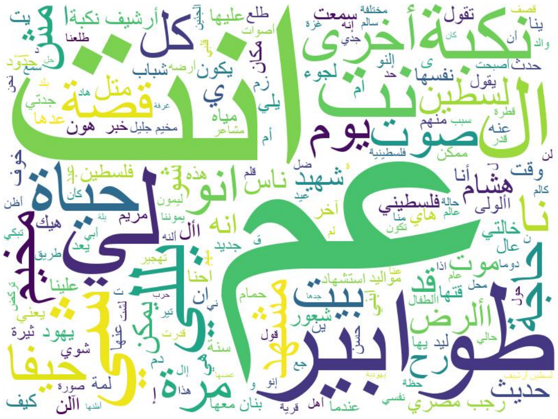

Figure 1: The Nakba stories presented as word-cloud.

in the dataset, with the size of each word corresponding to its frequency of occurrence. It can be clearly seen that the Arabic word طوابير /TwAbyr/, (Eng: Queues) is among the most frequent words in the Nakba stories. TwAbyr represents suffering in the daily life activities at refugee camps.

Being done with data handling steps, which are depicted using flowchart in Figure 2. We are ready for the next steps that relate to corpus construction.

4 Corpus Construction

This section presents our approach to developing a topic-classification dataset derived from Nakba narratives. Given that Arabic NLP research has historically received less attention than research on Western languages (Darwish et al., 2020), our work aims to enrich Arabic NLP resources with this specialized Nakba corpus.

4.1 Data Annotation

The topics covered in the Nakba stories were classified into five categories (including "Others"), as shown in Table 1. Annotating the corpus involved labeling the content of all stories, sentence by sentence, using one of the five categories. To achieve this, we employed a two-phase process that combined manual and automated annotation approaches.

In the first phase, a random sample of 33.33% of the dataset was manually annotated by two experts to establish a ground truth. This subset was carefully reviewed and refined over two iterative rounds, achieving an inter-annotator agreement (IAA) of 87%, which increased to 100% after resolving disagreements. The resulting ground truth subset served as the benchmark for evaluating various automated annotation methods.

In the second phase, the remaining two-thirds of the corpus were annotated using a rule-based classification system, which leveraged thematic

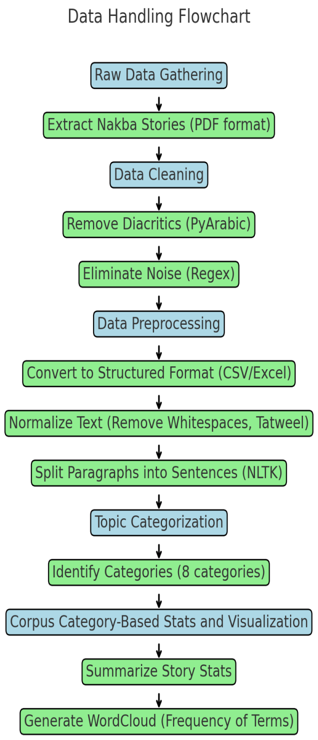
Data Handling Flowchart
Raw Data Gathering
Extract Nakba Stories (PDF format)
Data Cleaning
Remove Diacritics (PyArabic)
Eliminate Noise (Regex)
Data Preprocessing
Convert to Structured Format (CSV/Excel)
Normalize Text (Remove Whitespaces, Tatweel)
Split Paragraphs into Sentences (NLTK)
Topic Categorization
Identify Categories (8 categories)
Corpus Category-Based Stats and Visualization
Summarize Story Stats
Generate WordCloud (Frequency of Terms)

Figure 2: Data handling for Nakba stories.

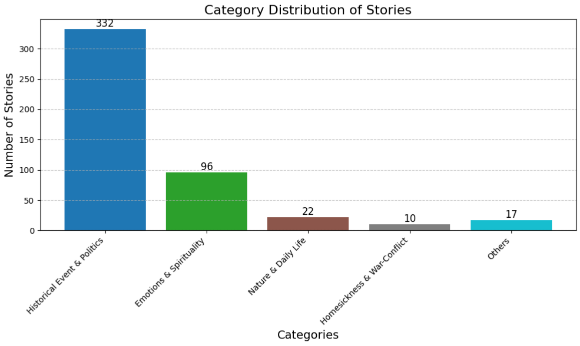

Figure 3: The category distributions in all stories.

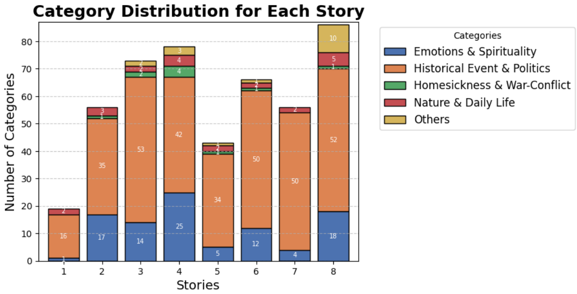

Figure 4: The category distributions for each story.

keywords and linguistic patterns identified from the manually annotated subset. For example, sentences mentioning terms such as *"diaspora"* or *"exile"* were categorized under *Homesickness & War/Conflict*, while sentences referencing political events were classified as *Historical Events & Politics*. This rule-based method, refined iteratively, demonstrated the highest accuracy compared to other methods such as KMeans clustering (Lloyd, 1982), TFIDF (Bafna et al., 2016), and AraBERT (Antoun et al., 2020).

The evaluation of these models focused on their ability to replicate expert annotations without additional training or fine-tuning. The aim was not to replace expert annotations for this relatively small dataset of 470 sentences but to assess the feasibility of using these models as scalable tools for annotating larger datasets in the future. By combining manual annotations with the rule-based system, we ensured consistency and reproducibility, creating a reliable corpus for advancing Arabic NLP studies.

These results are obtained using the rule-based (RB) classifier. Figure 3 presents the distribution of stories across five categories: Historical Event & Politics (HEP), Emotions & Spirituality (ES), Nature & Daily Life (NDL), Homesickness & War-Conflict (HWC), and Others. The "HEP" category dominates with 332 entries, followed by "ES" with 96 entries. Categories such as "NDL," "HWC," and "Others" are represented by fewer entries, with 22, 10, and 17 entries, respectively.This distribution suggests a strong emphasis on political and emotional themes in the dataset, while other topics, such as daily life and conflict-related stories, are less prevalent. The imbalance in category distribution points to the dataset's thematic concentration on historical and political narratives, highlighting potential gaps in diversity regarding more personal or nature-related themes.

Figure 4 illustrates the distribution of categories across eight stories, as obtained using the rule-based classifier. "HEP" appears most frequently, followed by "HWC." "ES" and "NDL" are represented less frequently, while the "Others" category is rarely mentioned. Story 8 stands out with the highest number of categories, with "HEP" and "HWC" being the dominant themes. This chart highlights the varying thematic focus of the stories, with political and emotional topics being more prevalent in the dataset.

4.2 Inter-Annotator Agreement

It is important to review the resulting topic classification labels obtained using the different three approaches. With assistance from two experts, the

ID	Type/ Approaches	Arabic Text
195	Agreement/ Expert 1 vs. Expert 2	إنه ملكي، إنها حريّتي وهذه أرضي
332	Disagreement/ Expert 1 vs. Expert 2	فلسطيني لبناني المهم ما في فلسطيني سادة
341	Agreement/ RB vs. Experts	لم أ صدق أن الله كرمني وجعلني ألتقي بأم شهيد
361	Disagreement/ RB vs. Experts	يا ريتني قدرت

Table 3: Examples of agreement and disagreement encountered during annotation.

inclusion of human factor allows not only for measuring Inter-Annotator Agreement (IAA), but also for helping to quantify the subjectivity of the task and refine the category definitions.

To assess the reliability and consistency of the three approaches, we calculated the IAA using Cohen's Kappa (Cohen, 1960) as shown in Equation 1. This follows Artstein and Poesio (2008), who applied Cohen's Kappa in computational linguistics.

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{1}$$

where

$$\begin{aligned} p_o &= \text{the observed agreement,} \\ p_e &= \text{the expected agreement by chance.} \end{aligned}$$

We computed the IAA using Equation 1 and got an initial rate of 87%, which indicates substantial agreement. Discrepancies in annotation were often due to the complexity and interference of topic expressions. To resolve these differences, we organized a post-annotation step where annotators discussed and clarified difficult cases. Through these discussions, we updated our annotation guidelines and conducted a second round of annotation, yielding an improved Kappa score of 100%.

Table 3 shows examples of agreement and disagreement that were encountered during annotation. For example, the sentence 195 (Eng: It's mine, it's my freedom and this is my land) belongs to "Emotions & Spirituality" was annotated correctly by both experts. Whereas, the sentence 332 (Eng: Palestinian Lebanese, the important thing is that there are no pure Palestinians) belongs to "Historical Event & Politics" was annotated correctly by only one expert, the other annotated it as "Emotions & Spirituality". Another example – among rule-based (RB) and the experts, the sentence 341 (Eng: I could not believe that God had honored me and made me meet the mother of a martyr.) belongs to "Emotions & Spirituality" was annotated correctly by both the experts and RB. Whereas, the sentence 361 (Eng: I wish I could) belongs to "Emotions & Spirituality" was annotated correctly by the experts, and annotated wrongly by the RB, namely as "Historical Event & Politics".

4.3 Evaluation Metrics

The evaluation focused on the ability of different automated methods to replicate expert annotations without additional training or fine-tuning. The models tested included:

- **Rule-based system:** Using thematic keywords and contextual patterns derived from the ground truth data.
- **KMeans clustering:** Applied directly to the dataset to group similar sentences.
- **TFIDF:** Used to extract features for classification based on term importance.
- **AraBERT:** A pre-trained Arabic language model used for sentence classification.

The performance of each model was compared against the manually annotated ground truth. The rule-based system achieved the highest accuracy, demonstrating its effectiveness in capturing thematic categories. This evaluation highlights the potential of automated models for annotating large-scale datasets, even without further training or fine-tuning.

For this task, two experts were recruited to review and annotate a sample of 159 sentences (i.e. one third) of the data independently. As inspired by Artstein and Poesio (2008), this iterative process helped ensure that the final corpus annotations are consistent and reliable.

We are utilizing accuracy, calculated as in Equation 2, to evaluate the different classification approaches. The accuracy of each classification approach (on the one-third of the data) is calculated by comparing it to the manually annotated sample.

The confusion matrix (CM) in Figure 5 illustrates the performance of the rule-based approach in classifying Arabic text into five categories: Historical Event & Politics, Emotions & Spirituality, Homesickness & War-Conflict, Nature & Daily Life, and Others. The matrix shows correct and incorrect predictions per class. Notably, the rule-based model performs best in classifying Historical Event & Politics, with 59 correct predictions. However, it struggles in distinguishing between more similar categories, such as Emotions & Spirituality and Homesickness & War-Conflict, as evidenced by misclassifications.

When comparing this rule-based model to other approaches, such as TFIDF and AraBERT, the rule-based model outperforms them with a higher accuracy of 61.33%. In contrast, the TFIDF and AraBERT models achieve significantly lower accuracy rates of 25% and 34%, respectively. This analysis demonstrates that the rule-based approach, despite its limitations in handling nuanced category distinctions, outperformed other methods by leveraging thematic keywords and contextual rules, as detailed in Section 4.1. In contrast, the lower accuracy of TFIDF and AraBERT underscores the challenges these methods face when applied to domain-specific datasets without fine-tuning.

This confusion matrix, therefore, represents the performance of the rule-based approach and highlights areas where improvement is needed, especially in distinguishing between certain categories. It is important to note that while the rule-based approach shows the highest accuracy, further research into hybrid models or the application of more sophisticated methods like AraBERT or TFIDF could lead to improvements in classification performance.

$$\text{Accuracy} = \frac{\text{Number of True Predictions}}{\text{Total Number of Predictions}} \quad (2)$$

4.4 Discussion

This study evaluated the feasibility of using automated models to annotate Arabic narratives, focusing on scalability for larger datasets. While manually annotating 470 sentences is straightforward, the aim was to test these models as tools for automating the annotation of thousands or millions of sentences in future research.

The rule-based classifier outperformed other approaches, leveraging thematic keywords and patterns to achieve the best accuracy. This suggests that carefully designed rule-based systems can effectively handle datasets with well-defined categories. However, pre-trained models like AraBERT exhibited lower accuracy due to the domain-specific nature of the Nakba narratives, highlighting the need for fine-tuning or specialized training for such contexts.

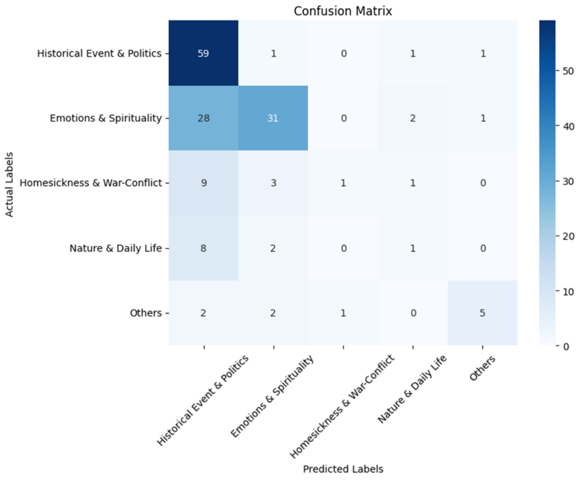

Figure 5: The confusion matrix (CM).

The low accuracy of ML models reflects Arabic NLP challenges, such as linguistic complexity and limited annotated datasets. The high inter-annotator agreement (87%, later 100%) highlights the reliability of manual annotations.

An essential aspect of this study was the grouping of categories, such as *"Emotions and Spirituality"* and *"Conflict and Homesickness."* This decision was informed by the natural co-occurrence of these themes in Nakba narratives. Emotional expressions are often intertwined with spiritual reflections, and narratives of conflict frequently evoke sentiments of homesickness. For instance, a sentence like *"My prayers keep me strong even as I endure exile"* captures both emotional and spiritual dimensions. Grouping these categories simplifies annotation, reduces ambiguity, and enhances the dataset's ability to capture intertwined themes.

However, this approach introduces limitations, as some sentences may lean more toward one aspect of a paired category. Future work should explore multi-label annotation schemes to better reflect the nuanced overlap between themes and provide more precise annotations.

These findings underscore the challenges of classifying Nakba narratives with existing models and emphasize the importance of expanding the dataset to address category imbalances. Future research should focus on fine-tuning transformer models and exploring advanced annotation schemes to overcome the limitations of small and specialized datasets. These enhancements will enable more robust analyses of Nakba narratives and contribute to advancing Arabic NLP research.

5 Conclusion

We presented the Nakba Topic Classification Corpus[2], an Arabic annotated dataset developed to support research in Arabic NLP, particularly in topic classification and emotion detection. Through careful preprocessing, annotation, and validation, we achieved good topic labels across five categories. This corpus is expected to bridge gaps in Arabic NLP resources and provide a foundation for future applications, including sentiment analysis and other machine learning classification tasks. By preserving and categorizing Nakba narratives, our work not only contributes to advancing Arabic NLP, but also serves as a vital resource for preserving and analyzing cultural narratives, offering a more in-depth understanding of the Nakba's historical and emotional dimensions.

Future research could expand the corpus by adding Nakba stories from diverse regions, Arabic dialects, and leveraging pre-trained models like GPT.

Acknowledgments

The authors would like to thank *Palestine Technical University - Kadoorie* in Tulkarm, Palestine for providing support.

References

Jeelani Ahmed and Muqeem Ahmed. 2021. Online news classification using machine learning techniques. *IIUM Engineering Journal*, 22(2):210–225.

Sallam AL-Sarayreh, Azza Mohamed, and Khaled Shaalan. 2023. Challenges and solutions for arabic natural language processing in social media. In *International conference on Variability of the Sun and sun-like stars: from asteroseismology to space weather*, pages 293–302. Springer.

Doha Albared, Hadi Hamoud, and Fadi Zaraket. 2023. Arabic topic classification in the generative and automl era. In *Proceedings of ArabicNLP 2023*. Association for Computational Linguistics.

Zaid Alyafeai, Mona Alshahrani, Maram Alenazi, Hesham Altwaijry, and Tayyaba Iqbal. 2022. A survey on arabic nlp: Where we are and where we are heading. *Journal of King Saud University-Computer and Information Sciences*, 34(6):2406–2423.

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

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Prafulla Bafna, Dhanya Pramod, and Anagha Vaidya. 2016. Document clustering: Tf-idf approach. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pages 61–66. IEEE.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

José Cabezas, Thomas Loiseau, and Juan Villena-Román. 2022. Cultural heritage preservation in the digital age: Challenges and approaches in natural language processing. *Journal of Cultural Heritage*, 57:105–118.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Kareem Darwish, Hamdy Mubarak, Ahmed Abdelali, Mohamed Eldesouki, and Sabit Hassan. 2020. A panoramic survey of natural language processing in the arab world. *arXiv preprint arXiv:2011.12631*.

Ashraf Elnagar, Ridhwan Al-Debsi, and Omar Einea. 2020. Arabic text classification using deep learning models. *Information Processing Management*, 57(1):102121.

Stuart Lloyd. 1982. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137.

Motaz K Saad and Wesam Ashour. 2010. Osac: Open source arabic corpora. In *6th ArchEng Int. Symposiums, EEECS*, volume 10, page 55.

Ahlam Wahdan, Mostafa Al-Emran, and Khaled Shaalan. 2024. A systematic review of arabic text classification: areas, applications, and future directions. *Soft Computing*, 28(2):1545–1566.

Taha Zerrouki. 2023. Pyarabic: A python package for arabic text. *Journal of Open Source Software*, 8(84):4886.

[2] https://github.com/PsArNLP/Nakba