Exploring Author Style in Nakba Short Stories: A Comparative Study of Transformer-Based Models

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

Abstract

Measuring semantic similarity and analyzing authorial style are fundamental tasks in Natural Language Processing (NLP), with applications in text classification, cultural analysis, and literary studies. This paper investigates the semantic similarity and stylistic features of Nakba short stories, a key component of Palestinian literature, using transformer-based models, AraBERT, BERT, and RoBERTa. The models effectively capture nuanced linguistic structures, cultural contexts, and stylistic variations in Arabic narratives, outperforming the traditional TF-IDF baseline. By comparing stories of similar length, we minimize biases and ensure a fair evaluation of both semantic and stylistic relationships. Experimental results indicate that RoBERTa achieves slightly higher performance, highlighting its ability to distinguish subtle stylistic patterns. This study demonstrates the potential of AI-driven tools to provide more in-depth insights into Arabic literature, and contributes to the systematic analysis of both semantic and stylistic elements in Nakba narratives.

1 Introduction

A very useful Natural Language Processing (NLP) tool is text similarity. It has many practical applications and allows us to find text that is similar to another text. The recent advancements in NLP have significantly improved the ability to analyze literary and cultural texts, offering new opportunities for exploring the richness of literary traditions. The development of deep learning models such as BERT and RoBERTa has made it possible to accurately capture the complexities of Arabic texts, a challenge that previous methodologies struggled to address (Devlin et al., 2019; Liu et al., 2019).

Nakba short stories, which depict the collective experiences and historical trauma of the Palestinian people, are an essential component of Palestinian literature and culture. However, analyzing Arabic literary texts computationally remains challenging due to the rich morphology, varied dialects, and cultural nuances of the language (Elmadany et al., 2020). Recent studies have shown the efficacy of transformer-based models in capturing these linguistic complexities, making them invaluable tools for semantic analysis (Antoun et al., 2020; Reimers and Gurevych, 2019). Furthermore, efforts to integrate cultural and historical context into text analysis have underscored the importance of domain-specific datasets and tailored pretraining strategies for effective semantic representation (Zaghouani et al., 2018; Saidi et al., 2024).

This research leverages deep learning models to measure semantic similarity and analyze authorial styles in Nakba narratives, aiming to uncover shared themes and stylistic variations among authors in this culturally significant genre. By applying advanced computational techniques, we investigate how well modern NLP models can grasp the nuanced linguistic structures and cultural context embedded within Arabic literary works. This study contributes to a more profound understanding of how these narratives convey social and political messages and demonstrates the potential of AI-driven tools to objectively and systematically analyze Arabic literature. This work sets a foundation for future research in using AI for cultural and literary studies. This work presents a framework and a database to discover whether new Nakba stories are similar to existing ones. To benchmark our results, we utilize TF-IDF (Bafna et al., 2016) as a baseline method, allowing for comparative analysis with advanced transformer-based models.

This paper is organized as follows. In Section 2, we present prior and recent research on Arabic text similarity. Section 3 provides a comprehensive analysis of the dataset, our analysis is enriched with visualization. Section 4 presents the experimental settings, describes the proposed approach and discusses our reported results. In Section 5, we

Proceedings of the Proceedings of the 1st International Workshop on Nakba Narratives as Language Resources, pages 56–62
January 20, 2025. ©2025 Association for Computational Linguistics

conclude and point out ideas for future search.

2 Related Work

Semantic similarity measures the extent to which two pieces of text share meaning, a vital task in various NLP applications, such as text classification, information retrieval, and machine translation. Arabic, with its complex morphology and diverse genres, presents unique challenges for semantic similarity tasks. This literature review focuses on recent machine learning-based approaches for measuring semantic similarity in Arabic text, emphasizing studies by Antoun et al. (2020), Reimers and Gurevych (2019), Saidi et al. (2024) and Ismail et al. (2022) respectively.

Antoun et al. (2020) introduced AraBERT, a transformer model pre-trained on large Arabic corpora, which has shown effectiveness in various Arabic NLP tasks, including semantic similarity. Its domain-specific adaptations make it well-suited for handling the morphological complexity of Arabic text.

Author style detection has gained attention in NLP as a means of exploring linguistic patterns unique to individual authors. Transformer-based models such as BERT and RoBERTa have demonstrated their ability to capture stylistic nuances, as seen in studies such as Reimers and Gurevych (2019) that explore sentence embeddings for stylistic comparisons. These methods are particularly relevant for Nakba short stories, where narrative styles vary from author to author.

Saidi et al. (2024) introduced a hybrid model combining BERT and a Gated Recurrent Unit (GRU) network for capturing semantic similarity in Arabic texts. Their model exploits BERT's powerful contextual embeddings to represent words in their specific contexts, which are then passed through a GRU layer to capture sequential dependencies. The study highlights the model's efficacy in handling various Arabic genres, demonstrating significant improvements over traditional methods. However, the reliance on large annotated datasets for fine-tuning remains a limitation, particularly given the scarcity of genre-specific Arabic corpora.

Ismail et al. (2022) proposed a semantic-based similarity approach using pre-trained word embeddings and a deep neural network architecture. Their model integrates multiple linguistic features, including morphological and syntactic information, to enhance the understanding of semantic relationships between Arabic texts. This approach has shown promising results in capturing subtle semantic nuances, especially in formal and classical Arabic genres. Nevertheless, the model struggles with informal and dialectal variations, which are prevalent in contemporary Arabic literature.

3 Data

Our data represents Nakba stories, which were taken from the Nakba Archive website[1]. These narratives have been previously explored in the context of topic classification, as introduced by Hamed and Zaidkilani (2025), providing a complementary resource for analyzing Arabic texts.

3.1 Raw Data Gathering

From the project section in the Nakba Archive website, we extracted a collection of eight Nakba stories that 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 Nakba short stories contain English alphabets, special characters, URLs ... etc. The diacritics were removed using the PyArabic (Zerrouki, 2023) Python library[2]. We removed the remaining noise using regular expressions.

3.3 Data Preprocessing

We want to have a dataset suitable for NLP tasks. We did so using three main preprocessing steps:

Convert to a 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. Among others, this includes data labeling or annotation for specific NLP tasks such as Named Entity Recognition (NER) or Sentiment Analysis.

Text Normalization With the help of PyArabic (Zerrouki, 2023), we normalized the Arabic text by removing the extra white spaces and the Tatweel.

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

[2] https://pypi.org/project/PyArabic/

Paragraph to Sentences Each paragraph was split into sentences based on the appropriate sentence ending and using the **NLTK** Python library. As a result, we had a CSV/Excel file with about 470 rows. This will enable us to annotate the sentences easily, and add the corresponding labels in the future.

3.4 Data Stats and Visualization

Here, we shed the light on data stats and provide a sort of text visualization with the help of WordCloud library (Mueller, 2022).

Nakba Stories Stats We are working with a dataset that comprises eight Nakba short stories (473 sentences), totaling approximately 12,288 tokens, making it relatively small but sufficient for exploring semantic and stylistic features within a focused literary genre. Table 1 summarizes the stats of the eight Nakba stories. It is important to notice that the eight stories were written by different authors, making the dataset well-suited for analyzing authorial styles alongside semantic similarity.

Nakba Stories as WordCloud To produce a meaningful wordcloud, (i) we removed the Arabic stopwords using NLTK, and (ii) we removed the word suffixes and prefixes using PyArabic library. The WordCloud in Figure 1 provides a visual representation of the most frequent words/terms in the dataset, with the size of each word corresponding to its frequency of occurrence. As shown in the bottom-right-corner, it can be clearly seen that the Arabic word /TwAbyr/[3], also in Table 2, (Eng: Queues), and transliterated as /TwAbyr/ is among the most frequent words in the Nakba stories. TwAbyr are very common at refugee camps, and represent the suffering faced during the daily life activities. The dataset used in this study, which includes Nakba short stories and their associated preprocessing steps, is publicly available on GitHub [4].

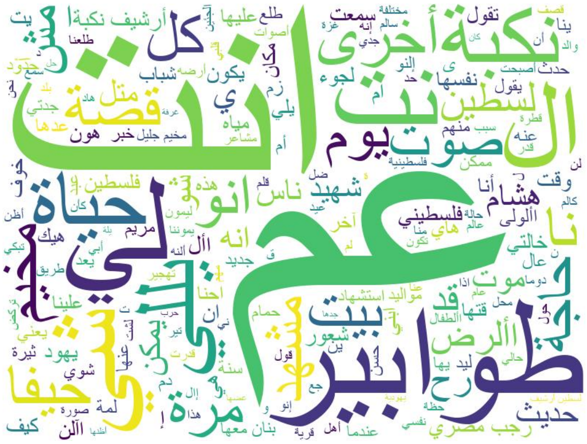

Figure 1: The Nakba stories presented as word-cloud.

4 Experimental Settings

In this study, we aim to measure semantic text similarity (STS) and analyze authorial styles in Arabic, focusing on Nakba short stories. Semantic similarity plays a crucial role in various NLP applications, such as text classification, information retrieval, and sentiment analysis, where understanding nuanced linguistic relationships is essential (Ismail et al., 2022). Recent advancements in transformer-based models like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have improved the ability to capture complex linguistic patterns, especially in underrepresented languages like Arabic. Our methodology leverages these pre-trained models to evaluate the semantic similarity across Nakba narratives, aiming to uncover recurring themes and patterns embedded within the cultural context of the stories. In addition to semantic similarity, we evaluate authorial style by analyzing linguistic features such as sentence complexity, word usage patterns, and lexical diversity. These features are extracted using embeddings from transformer-based models, enabling a detailed comparison of stylistic elements.

To benchmark these models, we use TF-IDF as a baseline, allowing for a comparative analysis of traditional versus deep learning approaches. TF-IDF was chosen as a baseline method for its traditional approach to measuring semantic similarity through term frequency and inverse document frequency. Despite its simplicity, TF-IDF provides a foundational perspective for evaluating the performance of more advanced models like BERT and RoBERTa.

4.1 System Description

BERT and RoBERTa were chosen due to their effectiveness in capturing intricate semantic relationships through deep bidirectional attention, an approach that excels in understanding context-sensitive word meanings (Devlin et al., 2019; Liu et al., 2019). This characteristic is particularly valuable for our study, as Nakba narratives often in-

[3] Broken Plural

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

ID	Arabic Name	English Name	Author	No. of Tokens
1	افتتاحية	Introduction	Ihab Kilani	556
2	ذاكرة لا تفنى	An Enduring Memory	Munira Al-Shehabi	1472
3	طوابير	Queues	Sarah Daoud	2082
4	أبو عادل الفاتح	Abu Adel, The Opener	Oula Jomaa	2069
5	حنين	Longing	Alaa Sukari	1113
6	الدخان فالخبز فالأسلاك الشائكة	Smoke, bread, & Barbed Wire	Ahmad Sukari	1479
7	أم الشهيد	The Martyr's Mother	Rama Abu Naaseh	1700
8	كيف لي أن أغفر	How Can I Forgive?	Shaimaa Taha	1817

Table 1: Biographies of the authors and statistics of the Nakba stories.

Arabic Word	Meaning	Transliteration	Singular /Transliteration/
طوابير	Queues	TwAbyr	طابور /TAbur/

Table 2: The Arabic word /TwAbyr/.

clude historical, emotional, and cultural nuances that simpler models might overlook. Additionally, BERT's and RoBERTa's subword tokenization approach is well-suited to handle Arabic morphology, which features root-based word variations and complex inflection patterns (Antoun et al., 2020).

NLP Pipeline Our NLP Pipeline aims to evaluate the semantic similarity between Arabic texts by employing State-of-the-Art (SotA) Arabic NLP techniques and machine learning models, specifically BERT and RoBERTa. This NLP pipeline encompasses the following steps as depicted in Figure 2.

- Data Preprocessing: Preprocessing is a critical step to prepare our dataset for optimal performance with these models. The normalization and preprocessing of Arabic was ensured in a previous step. Additionally, we employed Arabic-specific tokenization techniques to manage script and morphological complexity, utilizing subword units that allow the model to process word roots effectively.

- Feature Extraction: We utilized BERT-based and RoBERTa-based models to generate contextualized embeddings for Arabic texts. We also ensured that embeddings capture nuanced semantic relationships between words and phrases.

- Similarity Calculation: We employed cosine similarity as the primary metric to quantify the semantic similarity between texts. We also compared the embeddings generated by TF-IDF, BERT, and RoBERTa models to assess text similarity.

4.2 Experimental Results

We aim to explore the degree of semantic similarity and variations in authorial styles among the Nakba short stories. As inspired by (Cer et al., 2018; Reimers and Gurevych, 2019), we didn't compare all pairs of stories, instead we only compared the stories of similar lengths. Doing that is not only necessary to avoid biases, but also to increase accuracy.

As shown in Figure 3, the heatmap provides a detailed view of the similarity scores for story pairs, highlighting RoBERTa's superior performance across all pairs.

Table 3 presents the reported results for the four models. The TF-IDF serves as a baseline comparison, offering insight into the improvements achieved by transformer-based models in capturing semantic relationships within Nakba stories.

Both BERT and RoBERTa achieve high semantic similarity scores for Nakba story pairs, with RoBERTa slightly outperforming BERT due to its optimized pretraining. AraBERT, a transformer model specifically fine-tuned for Arabic, also demonstrates strong semantic similarity performance, achieving scores comparable to BERT and RoBERTa. This highlights the effectiveness of

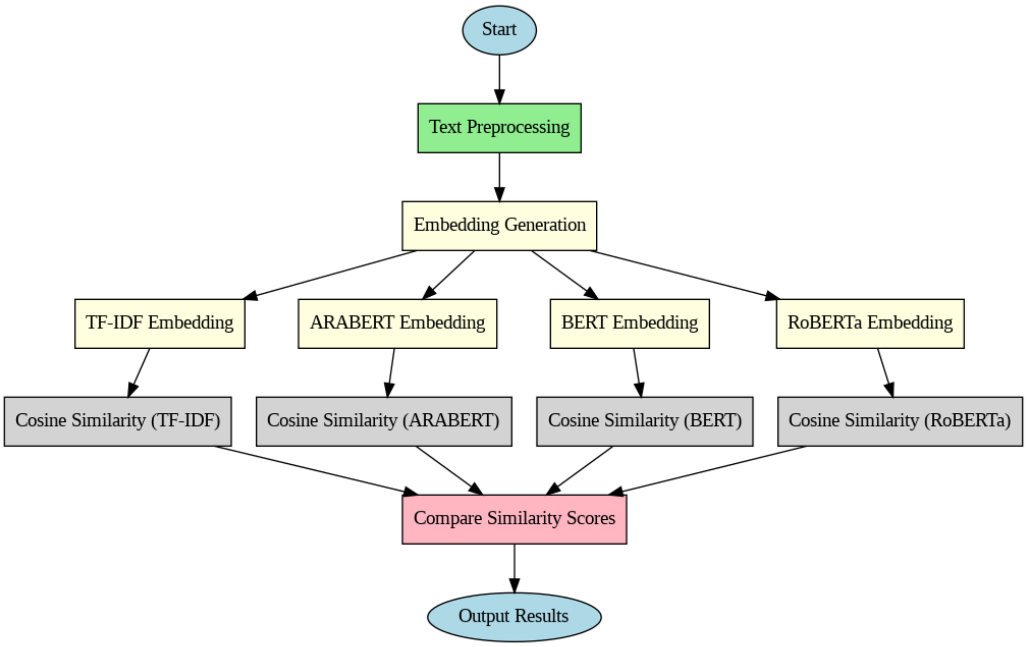

Figure 2: The Arabic text similarity analysis pipeline.

Criteria	Story 2 & 6	Story 7 & 8	Story 3 & 4
TF-IDF	54.40	42.16	**57.97**
AraBERT	**96.18**	93.87	95.23
BERT	**98.23**	97.89	97.77
RoBERTa	99.56	**99.77**	99.61

Table 3: Reported semantic similarity scores for Nakba short stories.

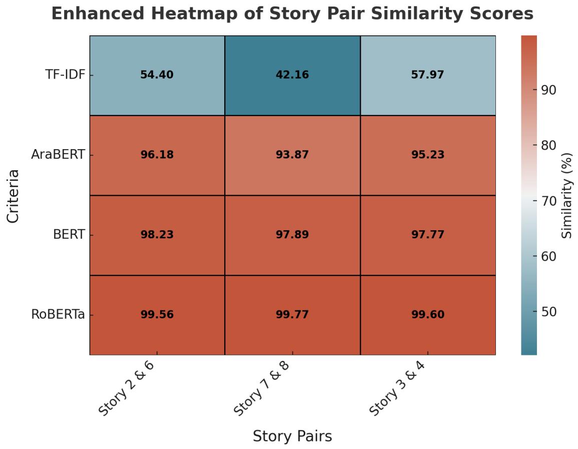

Figure 3: The heatmap of story pair similarities.

domain-specific adaptations in capturing Arabic linguistic and cultural nuances, despite slightly lower scores compared to RoBERTa. The results highlight shared cultural and historical themes, demonstrating the models' robustness. The results also indicate the ability of RoBERTa to capture subtle stylistic variations among authors, as reflected in differences in word choice and sentence construction. AraBERT, with its domain-specific pretraining, also performs well in identifying stylistic differences unique to Arabic.

4.3 Discussion

Table 3 reports the semantic similarity results for Nakba short stories using BERT and RoBERTa. Both models achieve high scores (above 97%), with RoBERTa slightly outperforming BERT, reflecting its superior contextual representation. Figure 4 illustrates the comparative performance of the models for each story pair, clearly demonstrating the gap between traditional TF-IDF and transformer-based models like BERT, RoBERTa, and AraBERT. These results highlight the models' ability to identify both shared cultural themes and distinct authorial styles in Nakba stories. For instance, RoBERTa excels in capturing stylistic variations, such as differences in sentence complexity and word usage patterns, which are critical for understanding individual narrative approaches. While the high scores demonstrate the models' robustness, they also high-

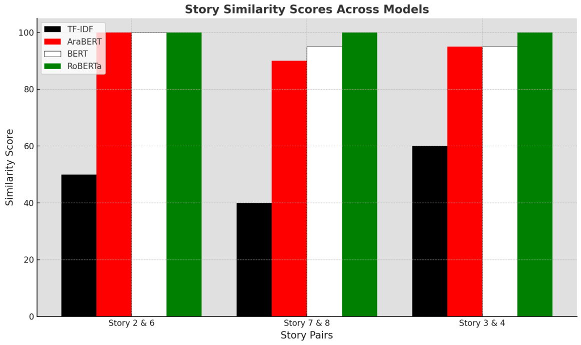

Figure 4: The stories similarities with different models.

light potential limitations in capturing subtle variations in stylistic or narrative details. Pairing stories of similar lengths ensures a fair evaluation by reducing bias.

The significant gap in performance between TF-IDF and transformer-based models like BERT and RoBERTa highlights the limitations of traditional approaches in capturing nuanced semantic relationships, particularly in linguistically complex narratives. These findings emphasize the potential of transformer models for broader literary analysis, including thematic clustering and pattern identification in Arabic texts.

5 Conclusion

In this paper, we investigated the semantic similarity and authorial styles of Nakba short stories using advanced transformer-based models, AraBERT, BERT, and RoBERTa. The high similarity scores achieved by all models underline their effectiveness in capturing nuanced linguistic structures, cultural contexts, and stylistic variations in Arabic texts. The slightly superior performance of RoBERTa highlights the impact of optimized pretraining on contextual representation and its ability to differentiate authorial styles within narratives. By comparing stories of similar length, we minimized bias and ensured a fair assessment of both semantic and stylistic relationships. This study demonstrates the potential of transformer-based models in the systematic analysis of Arabic literature, providing valuable insights into recurring themes, stylistic diversity, and shared cultural motifs.

Future work could focus on expanding the dataset to include a wider variety of Nakba narratives, allowing for more comprehensive analyses of semantic and stylistic variation. Incorporating additional stylistic features, such as syntactic complexity, lexical richness, and punctuation patterns, could provide more in-depth insights into authorial styles. Improving the models by fine-tuning on domain-specific corpora, adopting multitask learning approaches, or utilizing advanced architectures like SBERT for sentence-level embeddings could enhance their ability to capture subtle stylistic differences.

Acknowledgments

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

References

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*, pages 9–15.

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.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

AbdelRahim Elmadany, Hamdy Mubarak, Ahmed Kamal, Ahmed Eldin, Kareem Darwish, Tamer Elsayed, and Walid Magdy. 2020. Arabic offensive language detection with deep learning and text augmentation. In *Proceedings of the Fourth Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 37–42.

Osama Hamed and Nadeem Zaidkilani. 2025. Arabic topic classification corpus of the nakba short stories. In *Proceedings of the Nakba and Natural Language Processing Workshop (Nakba-NLP)*. To appear.

S. Ismail, M. Hassan, and W. Aref. 2022. Arabic semantic-based textual similarity. In *Proceedings of the Third International Conference on Arabic Computational Linguistics*, pages 65–72.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. In *arXiv preprint arXiv:1907.11692*.

Andreas Mueller. 2022. Word cloud: A little word cloud generator in python.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

A. Saidi, F. Al-Mousa, and H. Khalil. 2024. A bert-gru model for measuring the similarity of arabic text. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 102–110.

Wajdi Zaghouani, Nizar Habash, Behrang Mohit, and Houda Bouamor. 2018. Arabic nlp tools for the processing of arabic heritage in the digital age. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

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