# Arabic Sentiment Analysis: The Impact of Data Preprocessing on Model Performance

Ramadan Shemsu Hussen
Ortadoğu Araştırmaları Merkezi (ORSAM),Türkiye
Hacettepe University, Türkiye
Rhussen21@hacettepe.edu.tr

**Abstract**

This paper explores the role of data preprocessing in Arabic Sentiment Analysis (ASA) and its impact on model performance. The study was conducted as part of a broader research project on Arabic Natural Language Processing (NLP), specifically focusing on sentiment classification using deep learning models. My primary contribution, as a freshman researcher, was the preprocessing of Arabic and English text data to improve the quality of input for Convolutional Neural Networks (CNNs) used in the study. This paper presents the preprocessing pipeline, discusses its effect on dataset quality, and analyzes its role in enhancing sentiment classification accuracy. The original research project, *Empirical Evaluation of Word Representations on Arabic Sentiment Analysis*, demonstrated that CNN-based models trained on preprocessed data outperformed traditional machine learning approaches.

## 1 Introduction

Arabic Sentiment Analysis (ASA) is a complex NLP task due to the rich morphology of Arabic, its diverse dialects, and the informal nature of social media text. While deep learning models such as Convolutional Neural Networks (CNNs) have improved sentiment classification performance, the effectiveness of these models heavily depends on high-quality preprocessing.

This paper focuses on my role in the research project *Empirical Evaluation of Word Representations on Arabic Sentiment Analysis* [1], where I contributed to data preprocessing. The main goal of the project was to evaluate the impact of various unsupervised word representations on Arabic sentiment classification using CNNs. My responsibility was to clean, normalize, tokenize, and prepare the dataset before model training, ensuring the input text was suitable for deep learning-based sentiment analysis.

## 2 Background & Related Work

Previous studies on ASA have primarily relied on two approaches:

- **Traditional Machine Learning Methods** – Supervised classifiers such as Support Vector Machines (SVM) and Naïve Bayes have been widely used with handcrafted linguistic features like Part-of-Speech (POS) tags, sentiment lexicons, and morphological analysis [2].
- **Deep Learning Methods** – More recent research, including the main project, has leveraged CNNs trained on word embeddings (e.g., Glove, Skip-gram, CBOW) to achieve state-of-the-art performance [3, 4].

Preprocessing plays a critical role in ASA, as Arabic text contains diacritics, affixes, and stopwords that can mislead machine learning models. High-quality preprocessing enhances word embeddings and improves classification accuracy.

# 3 Methodology

## 3.1 Dataset Overview

The dataset comprised Arabic and English tweets used for sentiment analysis. The primary challenges included:

- Noisy text (misspellings, special characters, and redundant symbols).
- Mixed dialects and Modern Standard Arabic (MSA).
- Excessive stopwords affecting sentiment-bearing words.

## 3.2 Preprocessing Pipeline

As part of my role in the research project, I was responsible for designing and implementing a structured preprocessing pipeline to refine the dataset before training. The preprocessing stage was crucial in ensuring that the input data fed into the deep learning models, particularly Convolutional Neural Networks (CNNs), was clean, normalized, and representative of meaningful sentiment patterns. My supervisor emphasized the importance of understanding the relationship between raw text and model performance, allowing me to gain hands-on experience with real-world text processing challenges in sentiment analysis.

The preprocessing workflow I implemented focused on four key aspects: **text cleaning, tokenization and normalization, feature extraction, and network analysis.** These steps aligned with the broader research objective of evaluating word representations for Arabic Sentiment Analysis. By preparing high-quality input data, I contributed to the improved performance of the CNN-ASAWR model in the main study.

### 3.2.1 Text Cleaning and Standardization

The dataset contained tweets in both Arabic and English, requiring tailored cleaning approaches for each language. To ensure consistency in text representation, I:

- **Removed non-informative elements**, such as stopwords, punctuation, retweet indicators (RT), and special symbols like mentions (@) and hashtags, while preserving meaningful hashtag words.
- **Normalized text case and spacing** by converting all words to lowercase and removing redundant spaces, preventing duplicate word embeddings from being created for the same term in different cases.
- **Filtered out non-Arabic and non-English characters** to ensure that only relevant text was retained for analysis.

These steps reduced noise in the dataset, improving the quality of tokenized words used in word embeddings.

### 3.2.2 Tokenization and Linguistic Normalization

One of the key challenges in Arabic NLP is handling morphological variations and script inconsistencies. To address this, I applied:

- **Arabic word segmentation techniques** to break down words into their base components, improving the effectiveness of word embeddings.
- **Standardization of script variations**, such as normalizing different forms of hamza and unifying elongated or alternate spellings of common words.
- **Diacritic removal**, which was essential in reducing data sparsity, as diacritics are often inconsistently used in informal text.

These techniques helped create a more robust and uniform text representation, making it easier for the CNN model to extract meaningful sentiment features.

#### 3.2.3 Feature Extraction through N-gram Analysis and Visualization

To gain insights into the dataset's sentiment distribution, I performed exploratory analysis by:

- **Generating word clouds** to visualize the most frequently occurring words in positive, negative, and neutral tweets.
- **Extracting unigrams and bigrams**, identifying key words and phrases that contributed to sentiment classification.
- **Compiling statistical summaries of token frequency**, which helped assess dataset balance and potential biases in sentiment class distribution.

This analysis was instrumental in understanding the dataset's composition and informed later steps in feature selection for training the model.

#### 3.2.4 Network Analysis for Sentiment Clustering

To explore the impact of retweet activity on sentiment spread, I constructed a **retweet network graph** using Python and exported it to **Gephi** for visualization. This allowed us to:

- Identify **highly influential users** whose tweets significantly shaped sentiment trends.
- Observe how sentiment clusters formed within the dataset, offering insights into social media sentiment dynamics.

### 3.3 How This Contributed to the Research Project

The preprocessing pipeline I implemented played a foundational role in enhancing the quality of the dataset used in the main study. The CNN-ASAWR model relied on **pretrained word embeddings** such as Glove, Skip-gram, and CBOW, which perform best when trained on **clean, structured, and representative text.** My work ensured that the dataset met these conditions, ultimately contributing to the improved classification performance observed in the final research findings.

Additionally, my supervisor guided me through each stage to help me develop a deeper understanding of sentiment analysis and NLP methodologies. Through this experience, I not only gained proficiency in **data preprocessing, feature engineering, and visualization techniques**, but also learned how these techniques integrate into a broader machine learning pipeline. This hands-on exposure allowed me to build a strong foundation in **natural language processing**, setting the stage for future research contributions in Arabic sentiment analysis.

## 4 Conclusion and Future Work

This study highlights the importance of robust preprocessing in Arabic Sentiment Analysis. By implementing text cleaning, normalization, and tokenization, I improved dataset usability for deep learning models. My contributions ensured that the CNN-ASAWR model in the original project achieved superior results compared to traditional methods.

Future work could explore:

- Handling **dialectal variations** more effectively.
- Comparing different **lemmatization techniques** to improve feature extraction.
- Assessing the direct impact of each preprocessing step on **classification accuracy**.

## References

[1] M. Gridach, H. Haddad, and H. Mulki, *Empirical Evaluation of Word Representations on Arabic Sentiment Analysis*, Lecture Notes in Computer Science, Springer, 2017.

[2] B. Pang, L. Lee, *Opinion mining and sentiment analysis*, Foundations and Trends in Information Retrieval, 2008.

[3] T. Mikolov, et al., *Distributed representations of words and phrases and their compositionality*, Advances in Neural Information Processing Systems, 2013.

[4] J. Pennington, et al., *Glove: Global Vectors for Word Representation*, EMNLP, 2014.