Improving Dialectal Text Classification via Attention-Based Stopword Extraction for Algerian Arabic

No Author Given

No Institute Given

Abstract In low-resource dialects like the Algerian dialect, traditional stopword lists are often incomplete and poorly suited to informal text such as social media content. To address this challenge, we present a novel method for constructing a stopword list tailored to dialectal Arabic by leveraging model interpretability techniques. Building on the self-attention mechanism of MARBERT, a transformer model trained on Arabic and its dialects, we identify low-utility words based on their attention patterns during sentiment classification. Specifically, we track words that attract disproportionately high attention in misclassified samples but low attention in correct predictions, as these may serve as sources of noise rather than informative features. Candidate stopwords are dynamically evaluated through a probabilistic removal process during training. Experiments on a manually annotated Algerian dialect dataset demonstrate that the attention-based stopword extraction improves classification accuracy over static stopword models. This approach offers a practical solution for building effective stopword lists in low-resource dialects.

Keywords: Stopword List · Text Preprocessing · Algerian dialect · Sentiment Analysis · Attention Mechanism.

1 Introduction

Text preprocessing is a crucial step in most Natural Language Processing (NLP) pipelines, often involving the removal of stopwords, common words such as conjunctions, prepositions, and auxiliary verbs that carry limited semantic value [1]. Effective stopword removal reduces noise, improves computational efficiency, and can enhance the performance of downstream tasks like text classification [2].

While numerous stopword lists exist for Modern Standard Arabic (MSA) [1], dialectal Arabic remains largely under-resourced. The Algerian dialect poses unique challenges due to its high linguistic variability, rich morphology, and frequent code-switching with French. These characteristics make standard MSA stopword lists insufficient, as they fail to capture the informal expressions and frequent words typical of Algerian dialect, especially in social media texts [3].

Despite growing interest in dialectal Arabic processing [2], efforts to develop robust, domain-specific stopword lists for Algerian Arabic are limited. Existing

lists are often manually created or based solely on frequency heuristics, lacking semantic context and adaptability to modern NLP models [4].

Classical stopword removal methods have limited capacity to identify semantically irrelevant words in context-sensitive tasks [5]. In dialectal contexts, frequent words may carry nuanced meanings—such as sarcasm or negation, depending on usage and domain [6,7]. Static stopword lists do not adapt to task-specific needs, like sentiment classification, where word importance varies with context.

To address these challenges for low-resource dialects, we propose a model-driven approach that leverages the attention mechanisms of transformer-based language models. Inspired by prior work using attention for stopword identification [5], we fine-tune MARBERT [8], a transformer pretrained on Arabic and dialects, on a sentiment classification task and analyze its attention scores. We identify tokens consistently receiving low attention in both correct and incorrect predictions as candidates for a context-aware, task-specific stopword list.

The resulting Algerian stopword list is publicly available at `https://osf.io/ez342/?view_only=bdc2ef75b1ba4fc991b92dd9132eec78`, filling a critical gap in Arabic NLP resources and supporting the development of more accurate language technologies for underrepresented dialects.

This paper is structured as follows: Section 2 reviews related work on Arabic and dialectal stopword development; Section 3 details our proposed stopword extraction framework; Section 4 describes the experimental setup; Section 5 analyzes findings; Section 6 concludes and outlines future directions.

2 Related Work

Stopword lists for MSA have been extensively studied, with several manually and statistically curated lists available [1]. However, most Arabic dialect stopword lists rely on frequency-based heuristics or translations from MSA resources, which often fail to capture semantic irrelevance in context or adapt to task-specific needs [7].

For instance, [9] generated an Egyptian Arabic stopword list from online social network data, demonstrating improved sentiment classification using Naïve Bayes and Decision Tree classifiers compared to MSA lists alone. Similarly, [10] compared general, corpus-based, and combined stopword lists for MSA, finding that combined lists performed best with BM25 weighting. A supervised method combining domain-independent and domain-specific corpora for context-aware stopword detection was proposed by [11].

Other efforts include [12], which released a multi-dialect stopword package based on the Twitter corpus Kawarith, and [13], which developed a Moroccan dialect stopword list through a pipeline involving preprocessing, dictionary construction, and word embedding clustering on data from Facebook, Twitter, and YouTube.

Beyond Arabic, related work in other low-resource languages provides useful insights. For instance, [14] constructed a Hausa stopword list using TF-IDF and

manual validation on 1.2 million tokens from four news sources. The final list of 76 stopwords improved prior coverage by 6%, aiding tasks like sentiment analysis and machine translation. This highlights the value of combining statistical methods with expert review for effective stopword identification in under-resourced settings.

More recently, [5] proposed an attention-based method that identifies low-importance tokens across model predictions to dynamically generate stopword lists.

Despite these advances, no prior work has focused specifically on generating stopword lists tailored for the Algerian dialect. Our study builds on these ideas by applying attention-based stopword extraction to this low-resource dialect. We analyze MARBERT's attention patterns during sentiment classification to dynamically identify contextually irrelevant tokens and assess their impact on classification performance.

3 Methods for Stopword List Generation

This section outlines the methodology for constructing a stopword list for the Algerian dialect. We describe the dataset and compare two approaches: a classical Term Frequency-Inverse Document Frequency (TF-IDF), based method and a model-driven strategy using attention scores from a transformer. Both aim to identify contextually irrelevant tokens to improve classification performance.

3.1 DataSet

We used the publicly available dataset from [15], comprising 45,000 manually annotated social media comments in the Algerian dialect, labeled as positive, negative, or neutral. Originally created for sentiment analysis, the dataset was repurposed here to support stopword list generation. The sentiment labels enabled analysis of attention patterns to identify words with minimal sentiment value. Prior to model training, comments were preprocessed by removing punctuation, special characters, emojis, and normalizing whitespace.

3.2 Classical Stopword List Creation

We applied the TF-IDF metric to the cleaned dataset to identify low-information words. Tokens with TF-IDF scores below a set threshold were selected as stopword candidates. These were then manually filtered to exclude frequent yet semantically meaningful words. The final list includes 265 refined tokens, encompassing both dialectal and MSA terms common in Algerian Arabic social media.

3.3 Attention-based Stopword Extraction

We leverage the attention mechanisms of the MARBERT to identify contextually uninformative tokens. After fine-tuning MARBERT for sentiment classification,

we extract attention scores for each token from the final transformer layer. Tokens are grouped based on classification outcome, correct or incorrect, and a stopword score is computed as:

$$\text{Score}(w) = \mathbb{E}[\text{Attention}_{incorrect}(w)] - \mathbb{E}[\text{Attention}_{correct}(w)]$$

Higher scores suggest a token may contribute to misclassification, while low, consistent attention indicates low semantic value. We apply a multi-stage filtering process:

- **Attention thresholds:** Tokens with scores between 0.0001 and 0.03 are retained.
- **Frequency:** Tokens must occur at least 3 times in the dataset.
- **Length:** Tokens shorter than 2 characters are excluded.
- **Sentiment filtering:** Tokens manually verified against a sentiment lexicon are removed to avoid discarding sentiment-bearing words.
- **Named Entity Recognition:** Tokens identified as named entities via CAMeL Tools NER [16] are excluded to preserve semantic meaning.

The resulting list includes 374 tokens, capturing frequent dialectal and MSA terms in Algerian Arabic social media. These stopwords, though common, typically lack sentiment content. Table 1 presents a selection of stopwords from our newly generated attention-based stopword list.

Table 1. Examples of Local Algerian Arabic Stopwords with Explanations

Stopword	Explanation
دوك	Means "now" or "in a moment".
زعما	Means "supposedly" or "let's say".
معليش	Means "it's okay" or "no problem".
راك	Means "you are".
وكي	Means "when" or "how".
تاع	Equivalent to "of" or a possessive marker.
خويا	Means "my brother".

4 Experiments and Results

To evaluate the effectiveness of our attention-based stopword list for Algerian dialect sentiment classification, we conducted controlled experiments measuring the impact of different stopword removal strategies on classification performance. We used both traditional and deep learning models for a comprehensive assessment.

All experiments utilized the Algerian Dialect Sentiment Corpus [15], comprising 45,000 manually annotated social media comments labeled as positive,

negative, or neutral. We applied standard preprocessing: lowercasing, punctuation removal, and standard tokenization. A stratified 90/10 train-test split was employed.

4.1 Stopword Removal Configurations

We evaluated three distinct configurations:

- Dataset used without removing any stopwords, serving as the baseline.
- Stopwords identified by low TF-IDF scores.
- Our proposed list derived from MARBERT attention analysis.

4.2 Classification Models

Two classifiers were employed:

- **Support Vector Machine (SVM):** Implemented with LinearSVC and a TF-IDF vectorizer. Class imbalance was addressed via class_weight='balanced'.
- **Long Short-Term Memory (LSTM):** A recurrent neural network trained on 300-dimensional FastText embeddings for 10 epochs with early stopping. The architecture consisted of an embedding layer (128-dim output), a single LSTM layer (64 units), dropout (rate 0.3), followed by two dense layers (ReLU and softmax activations). The model was optimized with Adam and trained using sparse categorical cross-entropy loss.cross-entropy loss function.

4.3 Evaluation Metrics

Given class imbalance, we report macro-averaged Precision, Recall, and F1-score, which compute per-class metrics independently before averaging, ensuring equal weighting regardless of class distribution.

4.4 Results

Table 2 summarizes the sentiment classification performance of the SVM model under different stopword removal strategies.

The Attention-Based stopword list yields the highest overall accuracy of 69%, outperforming both the TF-IDF-based list and the baseline (no stopword removal), which both achieve 68% accuracy. This improvement, though modest, indicates that removing stopwords identified via the attention mechanism helps the classifier better focus on informative tokens.

Macro-averaged Precision, Recall, and F1-score metrics also favor the Attention-Based method, ranging between 0.66 and 0.67, compared to 0.64 to 0.66 for TF-IDF and 0.64 to 0.65 without stopword removal. This reflects a more balanced performance across sentiment classes. The weighted F1-score follows the same

trend, with the Attention-Based approach achieving 0.69, confirming the benefit of stopword removal weighted by class support.

Class-wise analysis reveals that the Positive class is consistently well classified across all setups, with precision and recall highest for the Attention-Based list (Precision: 0.72, Recall: 0.85, F1: 0.78), demonstrating that removing attention-based stopwords does not degrade detection of positive sentiment.

The Negative class gains the most from attention-based removal, improving precision (0.74 vs. 0.73) and recall (0.64 vs. 0.63) over the other methods, resulting in a better F1-score (0.69 vs. 0.68). This suggests that the attention-driven stopword list helps SVM better discriminate negative sentiment tweets.

The Neutral class remains the most challenging, with relatively low precision (0.53–0.55), recall (0.44–0.47), and F1 (0.48–0.51) regardless of stopword removal. This aligns with common difficulties in neutral sentiment detection, where tweets often overlap semantically with positive or negative classes.

Table 2. SVM Classification Report with Different Stopword Removal Approaches

Stopword Method	**Class**	**Precision**	**Recall**	**F1-score**	**Support**
Attention-Based	Positive	0.72	0.85	0.78	1891
	Negative	0.74	0.64	0.69	1595
	Neutral	0.55	0.47	0.51	1007
	Accuracy		0.69 (4493 samples)		
	Macro avg	0.67	0.66	0.66	4493
	Weighted avg	0.69	0.69	0.69	4493
TF-IDF	Positive	0.70	0.85	0.77	1891
	Negative	0.73	0.63	0.68	1595
	Neutral	0.53	0.44	0.48	1007
	Accuracy		0.68 (4493 samples)		
	Macro avg	0.66	0.64	0.64	4493
	Weighted avg	0.67	0.68	0.67	4493
No Stopwords	Positive	0.71	0.85	0.77	1891
	Negative	0.73	0.64	0.68	1595
	Neutral	0.53	0.44	0.48	1007
	Accuracy		0.68 (4493 samples)		
	Macro avg	0.66	0.64	0.65	4493
	Weighted avg	0.68	0.68	0.68	4493

Table 3 shows the sentiment classification performance of the LSTM model across the three stopword removal configurations.

The LSTM achieves the highest accuracy of 78% using the Attention-Based stopword list, outperforming both the baseline with no stopword removal (77%) and the TF-IDF-based approach (76%). This confirms that the attention-driven

stopword removal enhances the model's ability to capture sentiment-relevant features.

The macro-averaged Precision, Recall, and F1-score are all highest for the Attention-Based method, consistently around 0.74–0.75, compared to 0.72–0.74 for the other methods. Weighted averages also favor the Attention-Based list, with F1-score at 0.77, indicating improved balanced performance across classes.

Class-wise, the Positive class is the easiest for the LSTM to detect, with precision and recall highest under the attention-based approach (Precision: 0.79, Recall: 0.84, F1: 0.81). Similarly, the Negative class benefits from attention-based stopword removal, showing improved precision (0.75 vs. 0.72–0.74) and recall (0.71 vs. 0.70), resulting in a better F1-score (0.73).

Unlike in the SVM results, the Neutral class shows substantial improvement with attention-based stopword removal (F1: 0.69), compared to 0.66 and 0.68 in the other two setups. This suggests that the LSTM better leverages the cleaner input after removing attention-identified stopwords to differentiate neutral sentiment.

Table 3. LSTM Classification Report with Different Stopword Removal Approaches

Stopword Method	**Class**	**Precision**	**Recall**	**F1-score**	**Support**
Attention-Based	Positive	0.79	0.84	0.81	1891
	Negative	0.75	0.71	0.73	1595
	Neutral	0.70	0.68	0.69	1007
	Accuracy		0.78 (4493 samples)		
	Macro avg	0.75	0.74	0.74	4493
	Weighted avg	0.77	0.78	0.77	4493
TF-IDF	Positive	0.77	0.83	0.80	1891
	Negative	0.72	0.70	0.71	1595
	Neutral	0.67	0.65	0.66	1007
	Accuracy		0.76 (4493 samples)		
	Macro avg	0.72	0.73	0.72	4493
	Weighted avg	0.75	0.76	0.75	4493
No Stopwords	Positive	0.78	0.82	0.80	1891
	Negative	0.74	0.72	0.73	1595
	Neutral	0.69	0.67	0.68	1007
	Accuracy		0.77 (4493 samples)		
	Macro avg	0.74	0.74	0.74	4493
	Weighted avg	0.77	0.77	0.77	4493

5 Discussion

The experimental results demonstrate that stopword removal strategies significantly impact sentiment classification performance on the Algerian dialect dataset. Both classical (SVM) and deep learning (LSTM) models benefit from stopword lists tailored through attention-based analysis, outperforming traditional TF-IDF-based removal and the baseline without stopword removal.

As shown in Figure 1, the Attention-Based method achieves the highest classification accuracy for both models: 69% for SVM (compared to 68% for TF-IDF and no removal) and 78% for LSTM (outperforming TF-IDF at 76% and no removal at 77%). While the accuracy improvement for SVM is modest, the gain for LSTM is more notable, confirming the benefit of attention-guided stopword elimination. Macro-averaged F1-scores, reflecting balanced class performance, also favor the Attention-Based method with 0.66 for SVM and 0.74 for LSTM, compared to ranges of 0.64–0.65 (SVM) and 0.72–0.74 (LSTM) for other methods. These findings suggest that attention-based stopword removal helps both models better capture sentiment-relevant features by reducing noise from non-informative tokens. This advantage is especially important given the challenges of dialectal Arabic, including code-switching, informal language, and lexical variation. By leveraging model-internal attention weights, the stopword list becomes task-specific and context-sensitive, leading to more effective preprocessing than generic frequency-based methods like TF-IDF. Nevertheless, the improvements, while consistent, are moderate, reflecting the inherent challenges of sentiment analysis in dialectal Arabic, which involves linguistic variation, code-switching, and informal language use.

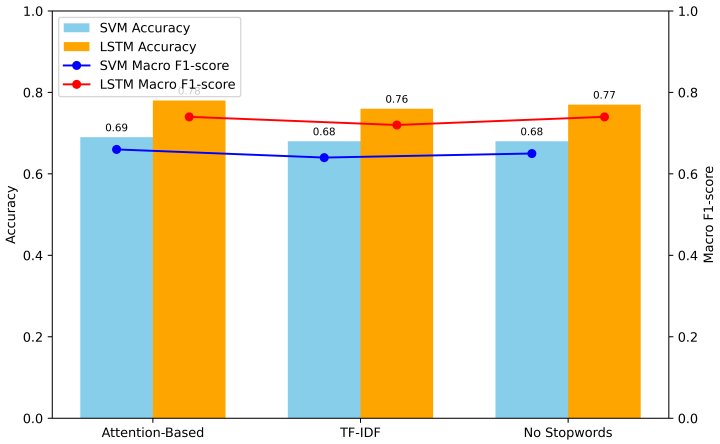

Figure 1. Classification performance comparison (Accuracy and Macro F1-score) of SVM and LSTM models with different stopword removal methods.

6 Conclusion and Future Work

We introduced an attention-based stopword removal method leveraging MARBERT to improve sentiment classification in Algerian Arabic. Experiments with both SVM and LSTM demonstrated improved accuracy and more balanced performance compared to TF-IDF-based and no stopword removal approaches. The attention-based method helped models focus on informative words, yielding better accuracy as well as more balanced precision and recall across sentiment classes, especially for challenging categories such as negative and neutral sentiments. These results confirm the potential of model-based preprocessing to better handle the complexities of dialectal Arabic text.

Future work will focus on further refining attention-based filtering and integrating it with advanced transformer architectures to enhance the robustness of sentiment analysis in low-resource dialectal settings.

References

1. Tengku Mohd Tengku Sembok and Belal Mustafa Abuata. Arabic stop words for information retrieval systems. *International Journal of Religion*, 6(1):121–127, 2025.
2. Yassir Matrane, Faouzia Benabbou, and Nawal Sael. A systematic literature review of arabic dialect sentiment analysis. *Journal of King Saud University - Computer and Information Sciences*, 35(6):101570, 2023.
3. Yassir Matrane, Faouzia Benabbou, Zineb Ellaky, and Chaimae Zaoui. *An Automatic Stop Words Removal in Maghrebi Arabic Dialect Text Classification Using Part of Speech Tagging*, page pp 187–196. 05 2025.
4. Zineb Nassr, Nawal Sael, and Faouzia Benabbou. Preprocessing arabic dialect for sentiment mining: State of art. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume XLIV-4/W3 of *ISPRS Archives*, pages 323–330. Copernicus Publications, 2020.
5. Yuki Kuwabara and Yu Suzuki. Automatic stopword generation based on attention for document classification using neural networks. *Journal of Information Processing*, 32:487–495, 05 2024.
6. Abdelrahman Kaseb and Mona Farouk. Saids: A novel approach for sentiment analysis informed of dialect and sarcasm, 2023.
7. Omar Alharbi. Negation handling in machine learning-based sentiment classification for colloquial arabic. *International Journal of Operations Research and Information Systems*, 11, 01 2020.
8. Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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, August 2021. Association for Computational Linguistics.
9. Walaa Medhat, Ahmed H. Yousef, and Hoda Korashy. Egyptian dialect stopword list generation from social network data, 2015.

10. Ibrahim Abu El-Khair. Effects of stop words elimination for arabic information retrieval: A comparative study, 2017.
11. D. Namly, K. Bouzoubaa, and A. Yousfi. A bi-technical analysis for arabic stopwords detection. *COMPUSOFT: An International Journal of Advanced Computer Technology*, 8(05):3126–3134, 2024.
12. Alaa Alharbi and Mark Lee. Kawarith: an Arabic Twitter corpus for crisis events. In Nizar Habash, Houda Bouamor, Hazem Hajj, Walid Magdy, Wajdi Zaghouani, Fethi Bougares, Nadi Tomeh, Ibrahim Abu Farha, and Samia Touileb, editors, *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 42–52, Kyiv, Ukraine (Virtual), April 2021. Association for Computational Linguistics.
13. Zineb NASSR, Nawal SAEL, and Faouzia BENABBOU. Generate a list of stop words in moroccan dialect from social network data using word embedding. In *2021 International Conference on Digital Age Technological Advances for Sustainable Development (ICDATA)*, pages 66–73, 2021.
14. Abubakar Salisu Bashir, Abdulkadir Abubakar Bichi, and Alhassan Adamu. Automatic construction of generic hausa language stop words list using term frequency-inverse document frequency. *Journal of Electrical Systems and Information Technology*, 11(1):58, 2024.
15. Zakaria Benmounah, Abdennour Boulesnane, Abdeladim Fadheli, and Mustapha Khial. Sentiment analysis on algerian dialect with transformers. *Applied Sciences*, 13(20), 2023.
16. Abdullah Aldumaykhi, Saad Otai, and Abdulkareem Alsudais. Comparing open arabic named entity recognition tools, 2022.