STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–21.
Published online in International Academic Press (www.IAPress.org)

Unveiling Quality-Factor Metrics to Optimize Data Collection: A Comprehensive Framework for Arabic Sentiment Analysis

Zouheir Banou, Sanaa El Filali*, El Habib Benlahmar, Fatima-Zahra Alaoui

Faculty of Sciences Ben M'Sick, Hassan II University, Casablanca, Morocco

Abstract Our study introduces a novel quality-factor-driven approach to dataset evaluation, focusing on four key attributes: Class Distribution Index (CDI), Topic Distribution Index (TDI), Average Inverse Document Frequency (IDF), and dataset size. By systematically analyzing these factors, this research assesses their influence on model performance in Natural Language Processing (NLP), particularly in Arabic sentiment analysis. The findings reveal that CDI and TDI exhibit substantial impacts, with CDI showing a strong positive correlation with accuracy (0.5568) and F1-score (-0.7808), indicating that while class distribution imbalance might help the model achieve higher accuracy, it adversely impacts its F1-score, thus reducing the balance between precision and recall. TDI also negatively affects accuracy and F1-score (-0.2242 and 0.2031), underscoring the challenges of uneven text distribution across datasets.
In contrast, Average IDF and dataset size positively correlate with model performance, with Average IDF contributing 0.2670 to accuracy and 0.3207 to F1-score, highlighting the predictive power of rare terms within the dataset. Dataset size further enhances F1-score (0.3540), reaffirming that larger datasets support improved sentiment classification accuracy.
This study provides foundational insights into the effects of dataset quality on Arabic sentiment analysis, offering strategic directions for future research in underrepresented languages and advancing our understanding of data quality's implications in NLP.

Keywords Arabic language, Binary classification, Data collection, Quality factors, Sentiment analysis

DOI: 10.19139/soic-2310-5070-2467

1. Introduction

The quality of datasets is a critical, yet often overlooked, factor in determining the success of machine learning (ML) models, particularly in the domain of sentiment analysis, which aims to identify the sentiment expressed in a piece of text, such as positive, negative, or neutral, across a single or multiple languages [1]. While numerous studies have focused on developing sophisticated models for sentiment classification, far fewer have systematically evaluated the quality of the data feeding these models. This gap is especially pronounced in Arabic sentiment analysis, where datasets can vary significantly in terms of class balance, linguistic diversity, and size. Such variations can lead to discrepancies in model performance, making it difficult to draw consistent conclusions across studies.

Data quality directly influences the accuracy, robustness, and generalizability of machine learning models. Poorly constructed datasets—whether due to imbalanced classes, limited topic diversity, or a lack of sufficient examples—can introduce bias, overfitting, or underperformance in sentiment classification tasks. This is particularly concerning in applications such as stock market predictions, where public opinions extracted from online sources are used to forecast market movements [2]. As such, understanding

*Correspondence to: Sanaa El Filali (Email: sanaa.elfilali@etu.univh2c.ma)

ISSN 2310-5070 (online) ISSN 2311-004X (print)

and measuring data quality becomes essential to improving both the design of sentiment analysis systems and the interpretability of their results.

In text classification tasks, particularly sentiment analysis, research often emphasizes preprocessing techniques, feature extraction methods, and model performance, while neglecting the dataset as a critical factor influencing model efficacy. This oversight frequently results in benchmarks with limited generalizability and practical applicability. Despite this, new research has begun addressing this gap. For instance, the data quality evaluation method proposed by [3] relies on aggregating the obtained F1-scores of multiple models. However, this approach is heavily influenced by several factors, such as preprocessing, the nature of the feature extraction method, and the model used. In contrast, our research introduces a novel framework for evaluating the quality of a text classification dataset directly based on its intrinsic attributes. This approach yields new metrics that are not affected by external factors, such as those mentioned above, but are instead inherently linked to the dataset itself.

This study introduces a quality-factor driven approach to dataset evaluation in the context of Arabic sentiment analysis. We propose four key quality attributes that directly impact model performance: Class Distribution Index (CDI), which quantifies the balance of sentiment classes within a dataset; Text Distribution Index (TDI), which assesses the variety of topics covered; Average Inverse Document Frequency (IDF), indicating the importance of rare terms; and the overall dataset size, which affects model generalization capabilities. By analyzing these quality factors, we aim to provide a systematic evaluation framework that will enable researchers and practitioners to better understand the strengths and limitations of their data before training models.

Our study applies this evaluation framework to multiple Arabic sentiment datasets, providing empirical evidence on the correlation between dataset quality and model performance. We demonstrate how variations in these quality factors can lead to significant differences in accuracy and F1-score, offering new insights into the critical role data quality plays in Arabic sentiment analysis. Furthermore, we explore how imbalances or deficiencies in these factors can introduce bias, limit generalizability, or inflate reported performance metrics, underscoring the need for standardized quality assessments in future work.

This paper seeks to answer the following research questions:

- **What are the characteristics of a high-quality dataset for sentiment analysis?**
- **Which dataset quality factors have the most significant impact on model performance?**

In the following sections, we review previous studies on Arabic sentiment analysis and examine how data quality has been treated in the literature. We then outline our experimental methodology, detailing how we applied our quality-factor framework to Arabic datasets, and present our findings. Finally, we discuss the broader implications of dataset quality in sentiment analysis and propose future directions for research in this area.

2. Related Work

This section critically reviews prior studies on Arabic sentiment analysis, with a focus on the machine learning and deep learning models used, as well as the impact of quality factors such as class balance.

2.1. Machine Learning Models

Support Vector Machines (SVM) have been widely used for Arabic sentiment analysis due to their effectiveness in handling high-dimensional feature spaces. SVMs work by constructing a hyperplane that differentiates data points across distinct classes, often using a one-vs-one or one-vs-all strategy [4, 5]. For instance, [6] achieved an SVM accuracy of 87% and an F1-score of 93%. In another study [7], SVM outperformed other algorithms, including Multinomial Naive Bayes (MNB), Random Forest, and LSTM, with a performance of 91.3% on a dataset of over 110,000 product review comments.

Other studies confirm SVM's robustness. For example, [8] analyzed the LABR book review dataset, showing that SVM achieved a 91% accuracy on an imbalanced dataset, outperforming other classifiers. Similarly, [9] and [10] report high SVM performance, achieving accuracies of 90% and 90.3%, respectively, on datasets such as the Open Corpus for Arabic (OCA) [11] and other domain-specific datasets.

2.2. Deep Learning Models

While SVM and traditional ML algorithms dominate earlier works, the rise of deep learning has introduced more complex models, such as Long Short-Term Memory (LSTM) networks and transformers. These models are particularly effective for large datasets due to their ability to capture sequential dependencies in text.

For instance, [12] employed a bidirectional LSTM model on the LABR dataset, achieving an accuracy of 81%. More recently, transformer models like BERT and its Arabic variant, AraBERT, have been applied to sentiment analysis tasks. [13] fine-tuned AraBERT on a Twitter-based dataset, reaching a 68% F1-score. These advancements demonstrate the growing importance of deep learning models in handling the complexity and nuances of Arabic text.

2.3. Impact of Quality Factors

A critical issue in sentiment analysis, particularly in Arabic datasets, is class imbalance. Many datasets exhibit skewed distributions, where one sentiment class is overrepresented. This imbalance can significantly affect model performance. For example, [8] compared SVM performance on both balanced and imbalanced datasets, finding that SVM performed better with imbalanced data when trained on bigram features.

Studies have also explored the effect of dataset size and linguistic diversity on model performance. Stemming, for instance, is often used to reduce word forms to their base forms and improve performance on frequency-based models like TF-IDF. In [14], using the Farasa Stemmer on a Moroccan Dialect dataset improved SVM accuracy to 78%. However, stemming can introduce challenges, especially with Arabic dialects, as noted in [15], where stemming produced an accuracy of 89% on a Jordanian Twitter dataset but led to errors in dialect-specific word forms.

2.4. Summary of Related Work

In summary, SVM remains a dominant model in Arabic sentiment analysis, particularly when paired with frequency-based embeddings like TF-IDF. Deep learning models, such as LSTM and transformers, are gaining prominence, particularly for handling large and complex datasets. However, quality factors, including class balance, dataset size, and linguistic diversity, significantly impact model performance. While stemming can improve performance in some contexts, it also introduces challenges, particularly for dialect-specific texts.

Despite the growing number of sentiment analysis studies in Arabic NLP, most focus on improving model architectures or preprocessing pipelines. Far less attention has been devoted to assessing the quality of datasets themselves, especially in terms of class distribution, topic diversity, or lexical rarity. Recent works like [3] explored sarcasm dataset quality by aggregating performance metrics across models, but such approaches remain dependent on external training conditions. In contrast, our framework directly evaluates datasets based on intrinsic attributes, offering a more stable and model-independent assessment of quality. This shift is essential to advance reproducibility and fairness in Arabic sentiment analysis.

Recent advancements in Arabic sentiment analysis have been comprehensively surveyed by [16], who highlight the predominance of deep learning methods and identify key challenges in the field. [17] evaluate the efficacy of large language models like LLaMA, Mixtral, and Gemma, noting their potential in Arabic NLP tasks. Among the models, LLaMA demonstrated superior comprehension abilities for the Arabic language, outperforming Mixtral and Gemma in both tasks. Notably, in Arabic-to-English translation, LLaMA surpassed the transformer baseline by 4 BLEU points. However, despite these strengths, all three

Table 1. Research work experiments summary

REF	Content Source	Dataset Name	Dialect(s)	Samples per class			Model	N-grams	Embed.	Acc	F1
				POS	NEG	NEU					
[6]	Twitter					0	SVM	U		87%	93%
[7]	Product Reviews			31803	32129	46827	SVM	U	TFIDF	91.3%	
[8]	Book Reviews	LABR	MSA	42000	8000	0	SVM	B	TFIDF	91%	91%
[9]	Twitter		Saudi	2408	1792	0	SVM	T		90%	
[19]							SVM	U		75	
[10]	Movie Reviews	OCA	Various	3137	4881	0	BEOA	U	Binary	84%	
[20]	Booking.com	Booking.com Reviews		254165	48590	71017	Log. Reg	B	TFIDF		94%
							SVM	U	TFIDF	97%	
[11]	Movie Reviews	OCA	Various	3137	4881	0	SVM	U	TFIDF	91%	
[14]	Twitter	MSTD	Moroccan	866	2769	6378	Linear SVM	U+B+T	TFIDF	78%	
[15]	Twitter	AJGT	Jordanian	900	900	0	SVM	B	TFIDF	89%	88%
[21]	Twitter	MARSA	Saudi	17327	20751	18726	SVM	U	TFIDF	76%	76%
[22]	Twitter	ASTD	Egyptian	799	1684	6691	Log. Reg	U+B	TFIDF	69%	
							SVM	U+B+T	TFIDF		63%
[23]	Twitter			75774	75774	0	Ridge Reg	U+B			99%
							Adaboost	U		99%	
[12]	Book Reviews	LABR	MSA	42000	8000	0	BiLSTM	U		81%	81%
[13]	Twitter	ASAD	Arabic	15215	15267	64518	Arabert (Fine-tuned)	U	Arabert		68%

models underperformed compared to state-of-the-art models in most cases. Additionally, [18] demonstrate the adaptability of transformer-based models such as XLM-R to morphologically rich languages, including Arabic.

The analysis of sentiment categorization across various Arabic dialects and sources shows that SVM and Logistic Regression models, particularly when paired with TF-IDF embeddings, are successful at capturing the intricacies of Arabic sentiment. These models routinely attain excellent accuracy and F1 scores, demonstrating their robustness in a variety of contexts, including social media and formal reviews. The investigation of machine learning approaches, ranging from traditional models to advanced algorithms like as Ridge Regression and AdaBoost, as well as the use of Arabic-specific embeddings such as Arabert, demonstrates the field's dynamic evolution.

3. Methodology

To evaluate the performance of each model, we suggest an experimental workflow (Fig 1) where we evaluate the performance of each algorithm on different datasets. We conducted the same preprocessing steps for every experiment during the whole study, including stop-word removal and stemming using the ARLStemmer [24] to revert every word to its radical form.

3.1. Benchmark datasets

The datasets were selected based on specific criteria, including the ratio of negative to positive tweets and the overall size of the datasets. Since these datasets were not originally designed for binary sentiment classification, additional classes were removed to perform a Positive/Negative classification task. The choice of these datasets is justified by the diversity of dialects they represent, including Egyptian, Moroccan, MSA, and Gulf Arabic dialects. In our experiments, we selected these four datasets:

- **Arabic Sentiment Tweet Dataset (ASTD)** was created to classify tweets, gathered from Twitter [22]. It consists of 10,000 tweets originating from active Egyptian accounts, categorized into 799 positive tweets, 1684 negative tweets, 6691 objective tweets, and 832 mixed tweets.
- **Moroccan Sentiment Tweet Dataset (MSTD)** is a comprehensive multi-domain sentiment dataset in Moroccan dialectal Arabic is called the MSTD [14]. It was developed in response to the dearth of Moroccan datasets that are freely accessible for sentiment analysis applications. 8,848 tweets total from the MSTD are categorized into four sentiment groups: mixed, neutral, negative,

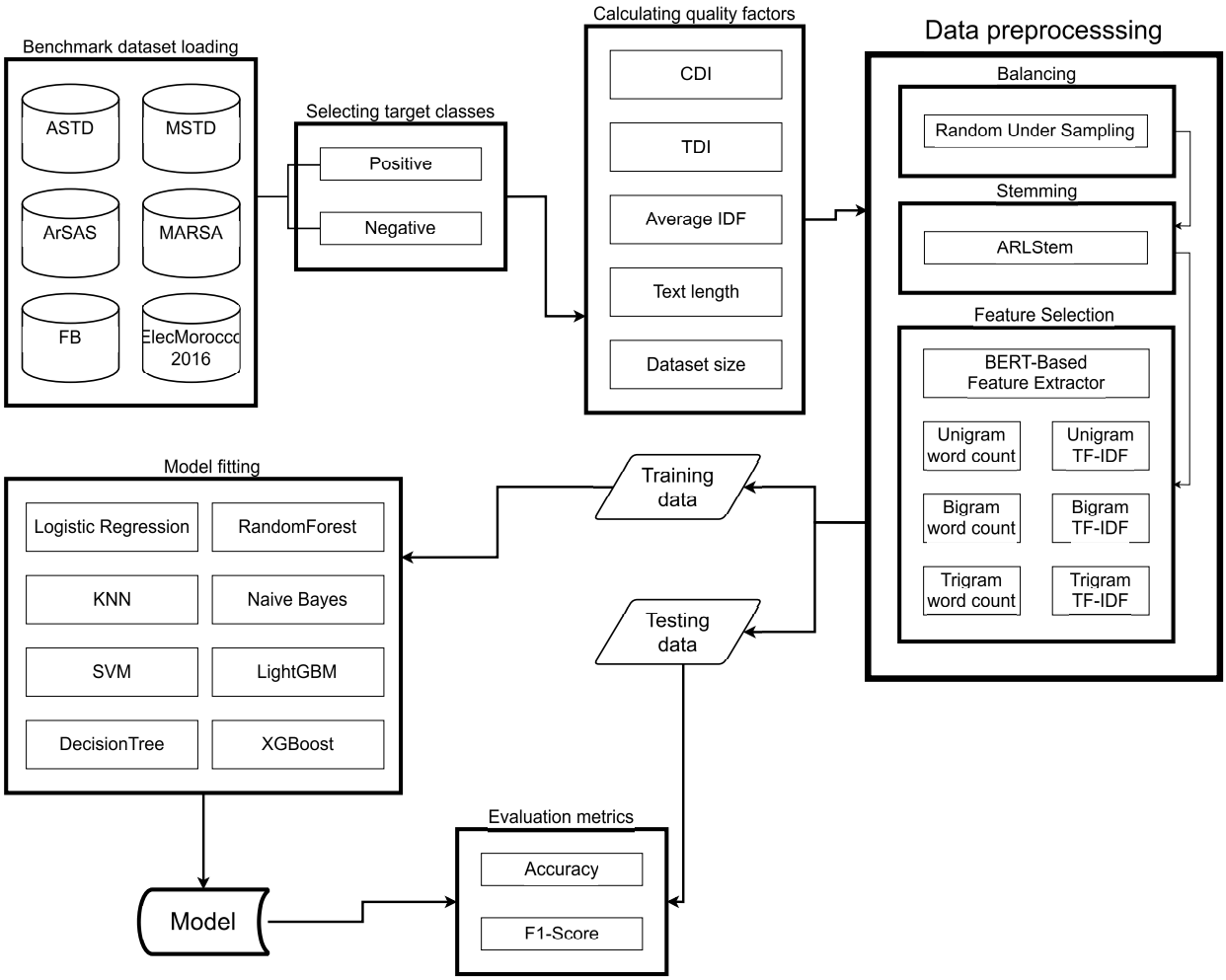

Figure 1. Experimental pipeline

and positive. The tweets touch on a variety of topics, such as social issues, sports, politics, and entertainment.

- **Arabic Speech-Act and Sentiment Corpus (ArSAS)** is a multipurpose dataset of Arabic tweets labeled for sentiment analysis and speech-act classification [25]. Most tweets in ArSAS are labeled as Negative (37.2%). Objective tweets are almost as frequent, constituting 34% of the dataset, whereas the rest of the dataset is split between Positive tweets (22%) and Mixed (6%).
- **Multi-domain Arabic Resources for Sentiment Analysis (MARSA)** is a corpus of tweets manually annotated for sentiment and categorized into four domains: social, political, sports and technology [21]. It is the largest resource of its kind for Gulf dialect Arabic.
- **Moroccan Facebook Comments Dataset (FB)** is a multipurpose dataset collected from Facebook comments, which goal is performing sentiment analysis in addition to the categorization of comments as written in Modern Standard Arabic or Moroccan Dialect Arabic [26]. For our experimental setup, we decided to only include the Moroccan Dialect subset.
- **Moroccan Elections-related dataset (ElecMorocco2016)** is a sentiment analysis dataset that revolves around a specific topic: the legislative elections in Morocco that took place in October 2016 [27]. Similarly to the FB dataset, only the Moroccan subset was included in our study.

The datasets used in this study exhibit various attributes in terms of class distribution (indicated by CDI and TDI indices), text length, average inverse document frequency (IDF), and overall dataset size. While some datasets are balanced (e.g. MARSA), others, such as ElecMorocco2016 and MSTD, show significant class imbalances (with CDI indices of 20.05 and 37.00, respectively), which may lead to decreased accuracy for underrepresented classes. These imbalances represent a limitation, as they can bias machine learning models toward the dominant classes. Resampling techniques, such as oversampling minority classes or undersampling majority classes, could be applied to mitigate this effect. Additionally, data augmentation techniques, such as generating synthetic text, could enhance the diversity of examples without altering the natural distribution of the text.

These datasets are well-suited for validating the study's results because they cover a broad range of distributions and sizes, allowing the assessment of how quality factors (CDI, TDI, Average IDF and Dataset Size) influence model performance. Their diversity in terms of quality and structure provides a robust framework for analyzing the effects of distribution indices and other attributes on sentiment classification performance in varied contexts. This justifies the use of resampling techniques to optimize performance in real-world applications.

3.2. Calculating Quality Factors

In sentiment analysis, especially when dealing with diverse text sources such as social media and product reviews, the quality of a dataset is a critical factor in determining the effectiveness of the predictive models. These quality factors include several key characteristics that influence model performance, such as the balance of sentiment classes and the linguistic diversity within the texts. Understanding these factors is crucial because they directly affect how well a machine learning model can learn from data and generalize to new, unseen data.

3.2.1. Overview of Quality Factors To systematically quantify dataset quality, we evaluate four key factors that affect sentiment analysis performance:

- **Class Distribution Index (CDI)**: Measures the balance between sentiment classes in the dataset.
- **Topic Distribution Index (TDI)**: Assesses the diversity of topics represented in the dataset.
- **Average Inverse Document Frequency (IDF)**: Captures the importance of rare terms within the text.
- **Dataset Size**: Determines the breadth of data available for training, impacting the model's ability to generalize.

3.2.2. Class Distribution Index (CDI) The **Class Distribution Index (CDI)** measures the balance of sentiment classes in a dataset. It helps determine if a model is trained equally across all classes, preventing bias toward one class. The CDI is calculated as follows:

Given a dataset D with N_C classes $\{C_i\}_{i=1}^{N_C}$, and p_i as the percentage of samples in class C_i, the mean sample size per class $\mu_C(D)$ is:

$$\mu_C(D) = \frac{1}{N_C} \sum_{i=1}^{N_C} p_i \tag{1}$$

The **CDI** is the standard deviation of class distributions, computed as:

$$cdi(D) = \sqrt{\frac{1}{N_C} \sum_{i=1}^{N_C} (p_i - \mu_C(D))^2} \tag{2}$$

The Class Distribution Index (CDI) measures how balanced the classes in a dataset are. A perfectly balanced dataset has a CDI of 0, indicating that all classes have equal representation. Conversely, a higher CDI reflects greater imbalance, with class proportions deviating significantly from the average class size.

For example, consider a dataset with three classes having proportions of 40%, 30%, and 30%. The average class size is calculated as 33.33% across all classes. The CDI reflects the aggregated deviation of each class's proportion from this average, providing a single numeric value that quantifies the overall balance of the dataset.

To make this metric more intuitive, we include a visual representation of the class distributions alongside the computed $\mu_C(D)$ value, as shown in Figure 2. The bar chart illustrates the percentage of samples in each class, with a horizontal dashed line indicating the average class size. The deviations of each class proportion from the mean are annotated above the corresponding bars *(+6.67% for Class 1, -3.33% for Classes 2 and 3)*. These deviations are aggregated to compute the Class Distribution Index (CDI), which quantifies the overall imbalance in the dataset.

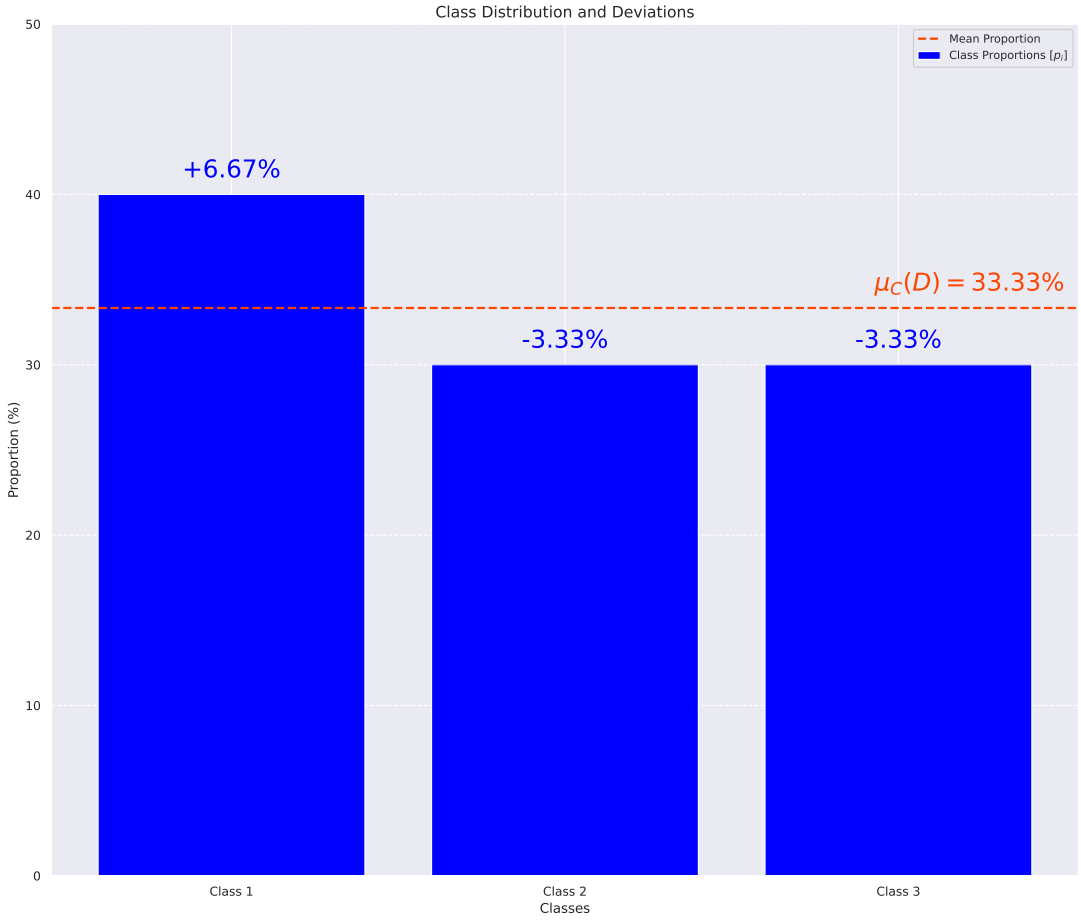

Figure 2. Visualizing Class Distribution and Deviations in a Dataset

3.2.3. Topic Distribution Index (TDI) The **Topic Distribution Index (TDI)** assesses the diversity of topics within the dataset. This helps determine whether the data covers a broad range of topics, which is crucial in applications like social media sentiment analysis. Given a dataset D with N_T topics $\{T_i\}_{i=1}^{N_T}$,

where t_i represents the percentage of topic T_i, the mean sample size per topic $\mu_T(D)$ is:

$$\mu_T(D) = \frac{1}{N_T} \sum_{i=1}^{N_T} t_i \tag{3}$$

The **TDI** is calculated similarly to CDI, as the standard deviation of topic distributions:

$$tdi(D) = \sqrt{\frac{1}{N_T} \sum_{i=1}^{N_T} (t_i - \mu_T(D))^2} \tag{4}$$

Here, $tdi(D)$ represents the degree of topic imbalance in the dataset. A lower TDI value corresponds to a more even distribution of topics, while a higher TDI indicates greater variability in topic representation across the dataset. It could also indicate that a dataset is more centered on a specific topic.

3.2.4. Average Inverse Document Frequency (ϕ) The **Average Inverse Document Frequency (ϕ)** reflects the importance of rare terms in distinguishing between sentiment classes. It is computed using the following formula:

$$\phi(D) = \frac{1}{|V_D|} \sum_{w \in V_D} \log\left(\frac{N}{n_w}\right) \tag{5}$$

Where:

- V_D is the set of all unique terms in the dataset.
- $|V_D|$ is the total number of unique terms.
- N is the total number of documents.
- n_w is the number of documents containing term w.

The average IDF measures the importance of rare terms within the dataset. Higher average IDF values indicate that the dataset contains more rare terms, which can be useful for distinguishing between sentiment classes.

Words with a high IDF value are rare but significant for sentiment classification, whereas common words with low IDF values contribute less to distinguishing between classes.

3.2.5. Dataset Size and Text Length The **Dataset Size** is another important quality factor. Larger datasets offer better representation of the language and sentiment distributions, enabling models to learn more robust patterns and generalize better to new data. Conversely, smaller datasets can lead to overfitting and lower generalization performance.

In frequency-based models, such as bag-of-words (BOW) and TF-IDF, **Text Length** indirectly influences model performance by affecting term frequency distributions. However, the dimensionality of the input vector remains fixed, regardless of text length. The variation in text length impacts how specific terms are weighted and represented within the document vector, subtly shaping the model's learning process.

3.2.6. Summary of Dataset Attributes We summarize these key quality factors and other dataset attributes in Table 2, which presents the characteristics of the benchmark datasets used in this study.

3.3. Data Preparation

Data preparation is a critical step in transforming raw data into a format suitable for machine learning models. This process involves cleaning and preprocessing the data to improve its quality and ensure compatibility with the selected algorithms. In this study, we applied several key preprocessing techniques to optimize the dataset for sentiment analysis.

Table 2. Benchmark dataset attributes*

Dataset	Samples (%)		Distribution indices (%)		Avg IDF	Text length	Dataset size
	POS	***NEG***	***CDI***	***TDI***			
ASTD Balanced	50.00	50.00	0.00	16.58	7.27	16	1256
ASTD	32.12	67.88	25.29	16.23	7.87	16	2419
ArSAS Balanced	50.00	50.00	0.00	16.93	8.30	22	3820
ArSAS	37.07	62.93	18.29	16.67	8.80	22	6429
ElecMorocco2016 Balanced	50.00	50.00	0.00	21.97	8.73	11	5906
ElecMorocco2016	35.82	64.18	20.05	21.75	9.24	12	10254
FB Balanced	50.00	50.00	0.00	6.83	7.14	15	1096
FB	19.31	80.69	43.40	7.13	8.26	15	3542
MARSA Balanced	50.00	50.00	0.00	10.97	10.06	13	27468
MARSA	45.37	54.63	6.55	11.07	10.43	13	37948
MSTD Balanced	50.00	50.00	0.00	8.43	7.36	15	1366
MSTD	23.84	76.16	37.00	8.76	8.30	16	3641

* Attributes concern datasets after filtering positive and negative tweets only

3.3.1. Balancing In machine learning, balancing refers to the process of correcting class imbalances in a dataset. When a dataset contains a disproportionate number of samples in one class compared to others, models tend to become biased toward the dominant class. To address this issue, we employed random under-sampling, which has proven to be more effective in our experiments compared to other balancing methods. This technique ensures that each class in the dataset is represented equally, thereby improving the model's performance and fairness. However, random undersampling may discard informative samples, potentially reducing model generalizability. In future iterations, we plan to explore alternative balancing techniques such as SMOTE and hybrid methods that retain class diversity while addressing imbalance.

3.3.2. Stemming Stemming is the process of reducing words to their root form, aiming to decrease the sparsity of input vectors, particularly in frequency-based embeddings such as TF-IDF or bag-of-words. While stemming can theoretically enhance model performance by grouping similar terms, its application poses challenges, particularly in languages like Arabic. The lack of universal stemming rules across languages may lead to incorrect modifications, especially in dialectal or foreign words.

For instance, as shown in Table 3, the Arabic word الوان (colors) is the plural of لون (color), but stemmers may mistakenly strip the definite article "الـ" as if it were the casing "ال". In 50% of cases, stemming tools may misinterpret this, altering the word incorrectly. Additionally, due to inconsistent spellings in social media posts (e.g., using "ا" instead of "ء"), stemming can further distort words.

Moreover, only 12.5% of the state-of-the-art studies reviewed in Section 2 achieved optimal performance through stemming. The remaining works either performed better without stemming or did not disclose its application. These findings highlight the error-prone nature of stemming in certain contexts, particularly for informal or dialectal texts, as exemplified in Table 3. A study by [28] has experimentally found similar results; better performances can be reached when stemming is not used. The same study found that lemmatization didn't yield better results either, which they explained by the lack of a stemming algorithm dedicated to dialectal Arabic. These results were corroborated by an experimental study conducted on 4 Arabic dialect benchmark datasets. With this study in consideration, and given that our work is conducted on Arabic dialect datasets as well, we did not consider word reduction techniques in order to preserve the word in its current form to avoid any potential information loss.

3.3.3. Root extraction Root extraction, which aims to reduce words to their root form rather than a simple stem, is particularly valuable in Arabic due to its root-based morphology. For example, the words

"كاتب" (writer), "كتابة" (writing), and "كتب" (books) all share the root "ك ت ب", and root extraction allows models to recognize this shared semantic basis. Root extraction can therefore improve the model's semantic understanding but is computationally intensive and may introduce complexity, especially with dialectal variations where the root structure may differ or be less clear.

3.3.4. Normalization Normalization is essential for handling inconsistencies in Arabic text, especially in informal sources like social media. Normalization often includes unifying letters with multiple forms (e.g., changing "ى" to "ي" and "ة" to "ه") and removing diacritics, which can reduce noise and increase data consistency. However, excessive normalization might remove subtle distinctions between words, impacting models' ability to capture sentiment or context nuances. Together, these techniques highlight the unique challenges in Arabic NLP; while they improve consistency and reduce sparsity, they can also risk oversimplifying complex linguistic features. Understanding these trade-offs is critical in balancing preprocessing depth with the need for nuanced model performance.

Table 3. Examples of stemming

Original	**ARLStem** [24]	**ARLStem2** [24]	**Cistem** [29]	**ISRI** [30]	**Lancaster** [31]	**Porter** [32]
سينما	سينم	سينم	سينما	سين	سينما	سينما
رومبوان	رومبو	رومبو	رومبوان	رومبو	رومبوان	رومبوان
اللاعب	لاعب	اعب	اللاعب	لعب	اللاعب	اللاعب
لاعب	لاعب	اعب	لاعب	لعب	لاعب	لاعب
الوان	وان	وان	الوان	وان	الوان	الوان
الالوان	الو	الو	الالوان	الو	الالوان	الالوان

3.4. Feature extraction

Given the scarcity of pre-trained language models for under-represented languages and dialects, frequency-based embeddings present a practical and accessible alternative for effective text representation. By relying on word frequency and distribution within the corpus, Frequency-based embeddings like Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) enable robust feature extraction without requiring large-scale annotated data. We include frequency-based embedding techniques in this study to explore their viability in such settings, providing insights into their potential as a solution when pre-trained models are unavailable or unsuitable.

In the process of converting textual data into vectors, we explored two frequency-based embedding techniques, namely Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF), in our experimental framework. This choice is driven by the fact that the primary difference between both techniques lies in the IDF component, which is one of the metrics considered in this study. Moreover, a systematic literarture review conducted by [33] shows that these two techniques tend to produce better results among all frequency-based techniques. The Bag of Words, implemented using Scikit-learn's CountVectorizer, transforms a given text into a vector that represents the occurrences of each word from the corpus within that text. As a result, the vector representation of a text becomes a sparse vector of integers. On the other hand, TF-IDF operates similarly by dividing the count vector by the length of the document, yielding the term frequency vector (TF). This TF vector is then multiplied by the inverse document frequency (IDF), creating a vector that reflects the frequency of a word within the corpus. Experiments encompassed various n-gram ranges, ranging from unigrams to trigrams. This approach aimed to capture negators within the same phrase as the words surrounding them. For instance, the sentence from MSTD, "ولد الاطلس شكون يقدر عليه الله يحفظك بعينه التي لا تنام", contains the negator "لا" followed by the verb "تنام" translating to "sleeps." Both words needed to be captured as a single expression, hence the utilization of n-grams. In addition to frequency-based methods, we also considered predictive approaches such as the utilization of BERT-based models in the current study. This exploration aimed

to assess the impact of employing a pre-trained language model on the overall performance compared to frequency-based methods.

3.5. Model Fitting

In our experimentation, we employed a diverse set of machine learning models, each varying in terms of linearity, probabilistic assumptions, and complexity of tree-based algorithms and ensemble methods. Specifically, we applied the following ten models: Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Tree, Random Forest, Bernoulli Naive Bayes, Multinomial Naive Bayes, Complement Naive Bayes, XGBoost, and LightGBM. The selection of these models was guided by their distinct approaches to handling classification tasks, with a focus on their ability to manage high-dimensional feature spaces, non-linear relationships, and imbalanced datasets.

3.5.1. Model Selection The rationale for choosing these models lies in their ability to capture different aspects of the data:

- **Logistic Regression**: Chosen for its simplicity and interpretability, especially in binary classification tasks, where it models the probability of class membership using a linear decision boundary.
- **K-Nearest Neighbors (KNN)**: A non-parametric model that assigns class labels based on the majority vote of the nearest neighbors, making it suitable for non-linear data.
- **Support Vector Machines (SVM)**: Used for its effectiveness in high-dimensional spaces, particularly in cases where classes are not linearly separable. The regularization parameter C, which controls the trade-off between margin maximization and classification errors, was tuned via grid search.
- **Decision Tree**: A tree-based model chosen for its ability to handle both categorical and continuous data, providing an interpretable structure through decision rules.
- **Random Forest**: An ensemble method that combines multiple decision trees to improve classification performance and reduce overfitting. The number of trees (`n_estimators`) and tree depth were optimized through grid search.
- **Bernoulli, Multinomial, and Complement Naive Bayes (NB)**: Probabilistic models selected for their efficiency in handling text data, particularly for datasets with categorical features or when features represent frequencies (e.g., TF-IDF). Each variant of Naive Bayes was selected based on its assumptions about feature distribution.
- **XGBoost and LightGBM**: Gradient boosting algorithms chosen for their robustness and ability to handle large datasets with minimal overfitting. Both models were tuned for optimal tree depth, learning rate, and number of boosting rounds.
- **GRU and LSTM**: Recurrent neural networks are typically used for processing sequential data, such as texts, and have shown reliable performance on different benchmarks, including Arabic sentiment analysis tasks [34]

3.5.2. Hyperparameter Tuning To ensure optimal performance, we employed grid search optimization with cross-validation to identify the best combination of hyperparameters for each model. The key hyperparameters considered for tuning are summarized in Table 5. Specifically:

- For **K-Nearest Neighbors (KNN)**, the number of neighbors k was varied from 1 to 19 to find the optimal neighborhood size.
- For **SVM**, the C parameter was varied between 0.01 and 1, and the degree of the polynomial kernel between $1, 2, 3, 4, 5$, optimizing for the balance between margin maximization and classification errors.
- For **Random Forest** we tuned the number of trees (`n_estimators`) from 1 to 39, and the maximum tree depth from 1 to 39, to control model complexity and prevent overfitting.

- For **XGBoost** and **LightGBM**, we optimized the boosting type (`gbdt, dart, goss`) and tree depth (1 to 39), as well as learning objectives such as logistic regression and binary logistic, to improve performance on imbalanced datasets.
- For **GRU** and **LSTM**, we explored both `unidirectional` and `bidirectional` to assess the effect of word succession as well as backward dependency.

Each model was evaluated using a grid of hyperparameter values, and the best-performing hyperparameter combinations were selected based on testing accuracy, F1-score, and overall model performance metrics.

Table 4. Algorithm properties

Model type	**Probabilistic**	**Tree-based**	**Linear**	**Ensemble**	**Recurrent**
Log. Reg.			X		
K Neighbors			X		
SVM			X		
Decision Tree		X			
Random Forest		X		X	
Bernoulli NB	X		X		
Multinomial NB	X		X		
Complement NB	X		X		
XGBoost		X		X	
LightGBM		X		X	
GRU			X		X
LSTM			X		X

Table 5. Hyper-parameter grid for each algorithm

Model type	**Hyperparameter**	**Values**
K Neighbors	neighbors	[1; 19]
SVM	C	[0.01; 1]
	degree	[1; 5]
Decision Tree	Max depth	[1; 39]
Random Forest	estimators	[1; 39]
	Max depth	[1; 39]
Bernoulli NB	alpha	[0.0; 1.5]
Multinomial NB		
Complement NB		
XGBoost	Max depth	[1; 39]
	objective	reg: logistic, binary: logistic
LightGBM	Boosting type	gbdt; dart; goss
	Max depth	[1; 39]
GRU	Bidirectional	true, false
LSTM		

4. Results

In this section, we delve into the impact that various factors—such as the dataset, embedding technique, and model characteristics—have on model efficacy. We encapsulate the aggregate scores across different

datasets, embeddings, and models. Metric scores garnered from our experimental analysis using BERT, Bag of Words (BOW) and TF-IDF are respectively consolidated in Tables 6, 7 and 8. To ensure the reliability and consistency of the model scores assessed in this study, we have employed average score values, thereby affirming that the results are not anomalous but indicative of the model's overall performance stability.

Table 6 presents the performance scores for models trained on BERT-based embeddings across various dataset conditions. It reveals that SVM excels with smaller, balanced datasets, while XGBoost is notably effective for imbalanced datasets. LGBM prefers smaller datasets, and LogReg generally performs well across other conditions. GRU and LSTM demonstrated fluctuating performance across datasets. While their scores were modest compared to tree-based models, they showed promise on well-balanced corpora such as MARSA and ASTD Balanced. Their behavior highlights the importance of sequential context, although they remain sensitive to data size and imbalance. This overview highlights the distinct advantages of each model in leveraging BERT embeddings to optimize sentiment analysis tasks.

Table 7 elucidates the performance metrics of models utilizing Bag of Words (BOW) embeddings, delineating distinct model proficiencies across varied dataset configurations. Notably, Multinomial Naive Bayes demonstrates pronounced efficacy in environments characterized by smaller, balanced datasets, a performance attribute paralleled by Light GBM when trained on BERT-based embeddings. This observation suggests a nuanced preference of Multinomial NB and Light GBM for similar dataset conditions albeit with differing embedding techniques. Additionally, XGBoost's capacity to adeptly manage imbalanced datasets emerges as a consistent advantage across both BERT and BOW embeddings, highlighting its robustness against class distribution challenges. Conversely, LogReg exhibits broad applicability, yielding commendable performance across a spectrum of conditions not specifically tailored to the aforementioned models. Through this comparative lens, the analysis underscores the strategic selection of models and embedding methodologies in optimizing sentiment analysis tasks, reflecting the intricate interplay between model capabilities and embedding preferences.

Table 8 offers a comprehensive analysis of model performances when trained on TF-IDF embeddings, contributing significantly to the discourse on sentiment analysis. The data reveal that Support Vector Machines (SVM) consistently favor smaller, balanced datasets, a trend that aligns with their performance under BERT embeddings, thereby underscoring a stable model preference across diverse embedding techniques. Additionally, Bayesian models, notably Complement Naive Bayes (CNB) and Multinomial Naive Bayes (MNB), are identified as particularly effective in processing datasets with elevated IDF values. This observation underscores their capacity to exploit the discriminative potential of less frequent, yet semantically rich terms. LogReg demonstrates pronounced efficacy within balanced dataset environments, mirroring its attributes observed with BOW embeddings. Contrarily, KNN manifests its unique capability to perform optimally within smaller, imbalanced dataset contexts, a characteristic not previously discerned in analyses utilizing BERT and BOW embeddings.

This comparative analysis, spanning Tables 6, 7, and 8, uncovers compelling patterns in model performance across various embeddings. Models such as Support Vector Machines (SVM) and LogReg display a consistent affinity towards certain dataset configurations regardless of the embedding technique employed, indicating a robustness in their application. In contrast, Bayesian models, including Complement Naive Bayes (CNB) and Multinomial Naive Bayes (MNB), along with KNeighbors, demonstrate specialized strengths that are closely associated with specific embedding methodologies. The consistent success of these models under particular conditions highlights the paramount importance of judicious embedding selection in enhancing model efficacy for sentiment analysis endeavors. Such strategic choices are especially crucial in the context of Arabic sentiment analysis, where the linguistic complexities necessitate a nuanced approach to model and embedding alignment.

Table 6. Average performance scores obtained on experimental benchmark datasets using BERT as embedding

Dataset	Metric	Log. Reg.	KNN	SVM	BNB	MNB	CNB	Dec. Tree	RF	XGB	Light GBM	GRU	LSTM
ASTD	A	80.21	79.67	83.73	78.8	80.89	80.62	78.08	82.34	86.34	86.31	66.12	66.12
	F1	75.36	68.83	75.74	75.19	77.83	77.58	73.29	75.99	82.51	82.06	66.12	66.12
ASTD Balanced	A	76.24	78.31	82.42	70.39	77.72	77.72	68.63	77.51	81.77	82.29	66.12	66.12
	F1	74.48	73.53	80.15	47.91	75.06	75.06	66.2	75.15	79.93	80.46	66.12	66.12
ArSAS	A	87.79	80.68	83.92	65.24	70.19	70.19	77.84	81.38	87.26	87.15	62.99	62.99
	F1	86.82	78.23	81.19	45.38	68.72	68.72	76.05	79.51	86.18	86.12	62.99	62.99
ArSAS Balanced	A	85.93	81.65	84.67	65.01	71.57	71.57	76.33	82.02	86.85	86.79	37.01	37.01
	F1	85.15	80.17	83.85	45.89	70.1	70.1	75.18	81.07	86.04	86	37.01	37.01
ElecMorocco 2016 Balanced	A	77.72	73.95	77.56	66.04	66.41	66.31	68.34	74.57	78.16	78	36.23	36.23
	F1	74.7	67.46	70.53	45.53	65.51	65.46	64.36	69.5	74.56	74.57	36.23	36.23
ElecMorocco 2016	A	75.38	71.8	75.37	66.17	66.16	66.16	65.03	70.59	75.23	74.84	65.09	64.31
	F1	73.77	68.67	73.35	45.72	65.31	65.31	63.22	68.75	73.47	73.13	65.09	64.31
FB	A	83.92	84.95	85.84	82.78	80	79.7	82.64	85.87	88.53	88.86	76.73	76.73
	F1	74.9	66.06	66.78	58.15	71.75	71.46	70.79	72.12	77.95	78.38	76.73	76.73
FB Balanced	A	73.62	80.04	81.81	82.11	80.25	80.25	67.72	74.88	78.85	79.4	26.66	76.73
	F1	67.49	69.77	73.47	56.54	71.75	71.75	61.18	67.55	72.57	73.06	26.66	76.73
MARSA	A	87.82	82.63	86.67	56.92	77.39	77.39	77.57	81.9	87.24	87	61.81	60.54
	F1	87.67	82.07	86.4	41.67	77.07	77.08	77.27	81.47	87.04	86.8	61.81	60.54
MARSA Balanced	A	87.74	82.89	86.83	57.06	77.32	77.32	76.12	81.96	87.19	86.73	45.86	45.86
	F1	87.63	82.52	86.65	42.41	77.01	77.01	75.92	81.72	87.05	86.58	45.86	45.86
MSTD	A	80.58	75.34	77.23	73.52	67.27	66.99	72.87	76.61	81	80.69	73.25	73.25
	F1	71.89	53.74	53.42	47.85	61.88	61.65	61.32	59.97	69.51	68.85	73.25	73.25
MSTD Balanced	A	73.25	71.9	76.58	74.35	66.55	66.55	63.37	70.89	75	75.57	73.39	49.66
	F1	68.92	63.61	69.38	47.35	61.25	61.25	58.46	65.26	69.96	70.73	73.39	49.66

Table 7. Average performance scores obtained on experimental benchmark datasets using BOW as embedding

Dataset	Metric	Log. Reg.	KNN	SVM	BNB	MNB	CNB	Dec. Tree	RF	XGB	Light GBM	GRU	LSTM
ASTD	A	77.62	41.83	69.75	66.79	60.55	57.79	72.86	69.73	74.42	69.25	72.93	72.31
	F1	67.85	38.8	43.05	51.77	58.35	55.9	56.85	43.27	62.44	51.69	72.93	72.31
ASTD Balanced	A	69.9	47.61	66.42	58.09	70.92	70.92	52.7	60.29	64.69	62.33	69.83	69.21
	F1	67.81	37.84	59.87	57.25	68.66	68.66	49.7	56.92	62.36	60.93	69.83	69.21
ArSAS	A	85.15	51.32	79.47	73.87	77.33	76.37	80.89	69.28	83.83	83.31	82.81	81.96
	F1	83.76	48.77	73.67	68.03	76.7	75.8	78.41	54.24	81.96	81.38	82.81	81.96
ArSAS Balanced	A	83.77	51.12	79.93	71.25	81.69	81.69	78.29	75.15	82.34	81.86	66.10	37.01
	F1	82.91	48.37	78.51	70.98	81.08	81.08	76.59	73.27	81.33	80.69	66.10	37.01
ElecMorocco 2016	A	79.29	66.81	73.63	70.17	69.28	67.1	74.28	68.03	77.79	76.06	78.69	78.69
	F1	76.18	63.49	63.06	62.98	68.3	66.29	67.23	49.14	73.93	70.92	78.69	78.69
ElecMorocco 2016 Balanced	A	76.81	62.64	74.12	64.73	74.5	74.5	74.15	71.65	74.95	74.33	41.83	36.52
	F1	75.46	59.67	69.49	64.36	73.68	73.68	68.62	66.38	73.24	71.87	41.83	36.52
FB	A	85.75	66.04	82.39	74.88	57.42	52.27	83.54	81.01	85.02	84.06	82.93	84.20
	F1	69.53	53.16	52.34	49.58	52.9	48.96	64.75	45.95	70.73	64.44	82.93	84.20
FB Balanced	A	79.55	50.05	75.42	42.37	72.36	72.36	73.58	68.41	71.79	72.23	80.54	85.47
	F1	72.07	45.65	65.1	41.06	67.2	67.2	64.71	59.72	64.39	62.44	80.54	85.47
MARSA	A	89.62	66.7	80.27	88.06	89	88.9	74.89	64.44	85.48	81.49	83.39	88.27
	F1	89.51	62.78	78.13	87.9	88.94	88.85	72.87	55.57	85.18	80.9	83.39	88.27
MARSA Balanced	A	89.36	63.25	81.32	87.17	89.13	89.13	74.32	71.96	85.85	81.93	54.18	52.65
	F1	89.28	53.88	80.73	87.15	89.07	89.07	72.56	69.25	85.68	81.55	54.18	52.65
MSTD	A	80.7	57.95	75.42	68.95	58.59	54.67	77.82	74.83	80.6	76.25	76.82	80.25
	F1	66.89	51.59	46.03	47.57	54.19	51.34	62.53	43.88	67.02	59.12	76.82	80.25
MSTD Balanced	A	71.33	47.44	70.06	50.88	71.32	71.32	61.89	64.25	64.93	63.96	73.25	79.29
	F1	67.1	40.62	62.49	49.76	67.55	67.55	55.91	58.23	61.32	59.52	73.25	79.29

Table 8. Average performance scores obtained on experimental benchmark datasets using TF-IDF as embedding

Dataset	Metric	Log. Reg.	KNN	SVM	BNB	MNB	CNB	Dec. Tree	RF	XGB	Light GBM	GRU	LSTM
ASTD	A	77.62	41.83	69.75	66.79	60.55	57.79	72.86	69.73	74.42	69.25	69.21	69.63
	F1	67.85	38.8	43.05	51.77	58.35	55.9	56.85	43.27	62.44	51.69	69.21	69.63
ASTD Balanced	A	69.9	47.61	66.42	58.09	70.92	70.92	52.7	60.29	64.69	62.33	67.56	64.67
	F1	67.81	37.84	59.87	57.25	68.66	68.66	49.7	56.92	62.36	60.93	67.56	64.67
ArSAS	A	85.15	51.32	79.47	73.87	77.33	76.37	80.89	69.28	83.83	83.31	83.28	83.44
	F1	83.76	48.77	73.67	68.03	76.7	75.8	78.41	54.24	81.96	81.38	83.28	83.44
ArSAS Balanced	A	83.77	51.12	79.93	71.25	81.69	81.69	78.29	75.15	82.34	81.86	74.81	73.41
	F1	82.91	48.37	78.51	70.98	81.08	81.08	76.59	73.27	81.33	80.69	74.81	73.41
ElecMorocco 2016	A	79.29	66.81	73.63	70.17	69.28	67.1	74.28	68.03	77.79	76.06	78.64	78.89
	F1	76.18	63.49	63.06	62.98	68.3	66.29	67.23	49.14	73.93	70.92	78.64	78.89
ElecMorocco 2016 Balanced	A	76.81	62.64	74.12	64.73	74.5	74.5	74.15	71.65	74.95	74.33	75.82	74.26
	F1	75.46	59.67	69.49	64.36	73.68	73.68	68.62	66.38	73.24	71.87	75.82	74.26
FB	A	85.75	66.04	82.39	74.88	57.42	52.27	83.54	81.01	85.02	84.06	82.93	84.49
	F1	69.53	53.16	52.34	49.58	52.9	48.96	64.75	45.95	70.73	64.44	82.93	84.49
FB Balanced	A	79.55	50.05	75.42	42.37	72.36	72.36	73.58	68.41	71.79	72.23	85.19	82.09
	F1	72.07	45.65	65.1	41.06	67.2	67.2	64.71	59.72	64.39	62.44	85.19	82.09
MARSA	A	89.62	66.7	80.27	88.06	89	88.9	74.89	64.44	85.48	81.49	88.25	88.29
	F1	89.51	62.78	78.13	87.9	88.94	88.85	72.87	55.57	85.18	80.9	88.25	88.29
MARSA Balanced	A	89.36	63.25	81.32	87.17	89.13	89.13	74.32	71.96	85.85	81.93	58.00	58.46
	F1	89.28	53.88	80.73	87.15	89.07	89.07	72.56	69.25	85.68	81.55	58.00	58.46
MSTD	A	80.7	57.95	75.42	68.95	58.59	54.67	77.82	74.83	80.6	76.25	79.29	73.25
	F1	66.89	51.59	46.03	47.57	54.19	51.34	62.53	43.88	67.02	59.12	79.29	73.25
MSTD Balanced	A	71.33	47.44	70.06	50.88	71.32	71.32	61.89	64.25	64.93	63.96	73.53	73.25
	F1	67.1	40.62	62.49	49.76	67.55	67.55	55.91	58.23	61.32	59.52	73.53	73.25

5. Discussion

5.1. Regression analysis

In this subsection, we applied Ordinary Least Squares (OLS) regression to examine the relationship between various quality factors and performance metrics, which served as our dependent variables in the context of sentiment classification. The OLS results provided key statistical outputs, including coefficients and p-values for each predictor variable, allowing us to assess their statistical significance in predicting performance metrics. These results are summarized in Table 9.

Table 9. Statistical metrics of quality factors in terms of each performance metric

Quality factors	**Coefficients**		**P-values**	
	Accuracy	***F1-score***	***Accuracy***	***F1-score***
CDI	0.00570	-0.51729	0.93937	0.00028
TDI	-0.66948	-0.36098	0.01923	0.17068
Average IDF	0.09915	0.07894	0.00001	0.00008
Text length	0.00198	0.00571	0.55103	0.13377
Dataset size	-0.28887	-0.17563	0.00108	0.01891

As shown in Table 9, both average IDF and dataset size were statistically significant across all performance metrics, indicating their strong influence on model outcomes. TDI (Topic Distribution Index), on the other hand, was only statistically significant with respect to accuracy, suggesting that topic diversity may play a role in how accurately a model classifies sentiments but does not impact other metrics as strongly. In contrast, CDI (Class Distribution Index) showed statistical importance in relation to the F1-score, highlighting its role in improving performance in imbalanced datasets. Notably, text length did not have a statistically significant impact on model performance, likely due to the nature of the feature extraction techniques (such as TF-IDF) used in this study, which tend to neutralize the influence of text length. To further interpret the relationships between these quality factors and model performance, we proceed with a correlation analysis in the following subsection.

5.2. Univariate analysis

In computational linguistics and machine learning, univariate analysis plays a crucial role in examining the distribution, central tendencies, and variabilities of individual variables. When applied to the performance analysis of sentiment classification models trained on datasets with specific attributes, univariate analysis helps identify which features contribute most significantly to enhancing model performance. By thoroughly analyzing these quality factors in isolation, we can better understand their individual impacts on the predictive modeling process, ultimately improving the interpretability and performance of the models.

Following this analysis, it essential to explore the correlations between these quality factors and the performance metrics—specifically accuracy and F1-score. Correlation coefficients indicate the strength and direction of the relationship between two variables, offering insight into how changes in one variable may be associated with variations in another. In the context of text classification, these coefficients between quality factors and performance metrics provide valuable information on the predictive validity of individual textual features. The correlation results are summarized in Table 10.

From Table 10, we observe a strong negative correlation between the F1-score and CDI (Class Distribution Index), indicating that as class imbalance increases, F1 performance decreases. In contrast, accuracy shows a positive correlation with CDI, suggesting that models tend to favor accuracy when the dataset is imbalanced, likely due to the dominance of a particular class. Conversely, TDI (Topic Distribution Index) exhibits a negative correlation with accuracy and a positive correlation with F1-score, suggesting that greater topic diversity challenges overall accuracy but enhances the model's ability

Table 10. Correlation coefficients between studied quality factors and performance metrics

Quality factor	Accuracy score	F1-score
CDI	0.5568	-0.7808
TDI	-0.2242	0.2031
Average IDF	0.2670	0.3207
Dataset size	0.296	0.3540

to balance precision and recall. Additionally, both average IDF and dataset size correlate positively with all studied metrics, with stronger correlations observed for F1-score, indicating their importance in improving the overall performance of the model, particularly in handling more complex or imbalanced data.

The behavior of CDI can be explained by the imbalance in the ratios of output labels within the dataset. As CDI increases, it reflects greater class imbalance in the training data, leading to models that are prone to overfitting. These models struggle to correctly classify samples from the minority class, resulting in a lower F1-score, which is sensitive to both precision and recall. However, accuracy remains high because it does not account for class distribution during its calculation; it focuses solely on the proportion of correct predictions, regardless of class imbalance. This explains why CDI correlates negatively with the F1-score and positively with accuracy. This result emphasizes how crucial it is to resolve class imbalance while building datasets and training models. Assessing the class distribution of the datasets has to be a top priority for practitioners, especially in fields where class imbalance is common. Techniques including data augmentation, loss weighting, and resampling (e.g., oversampling minority classes or undersampling majority classes) can be used to mitigate the adverse effects of high CDI. It is also advised that researchers use evaluation measures that explicitly account for class imbalance, such as the macro-averaged F1-score, rather than accuracy, which could mask the impacts of unbalanced data. Addressing class imbalance effectively enables models to achieve more robust and equitable performance across all classes.

TDI (Topic Distribution Index) reflects the diversity of topics within a dataset. A high TDI value indicates a dataset focused on a specific topic, while a lower TDI suggests a variety of topics are covered. In our study, we found that accuracy scores are statistically higher in datasets with a broad range of topics, whereas F1-scores tend to be higher in datasets concentrated around a single topic. This difference is influenced by the type of feature extractors used in our experiments, whether they were frequency-based or prediction-based.

To make CDI and TDI values more interpretable in real-world settings, we define qualitative thresholds:

- Values below 0.10 indicate **low imbalance**
- Scores between 0.10 and 0.20 reflect **moderate imbalance**
- Values above 0.20 signify **critical imbalance**.

These ranges help assess dataset readiness and guide practitioners in choosing appropriate preprocessing strategies.

In frequency-based embeddings, such as TF-IDF, datasets centered on a single topic often benefit from stronger feature indicators, like technical terms with high polarity. Similarly, in BERT-based embeddings, high-polarity terms are encoded as sets of robust features. However, datasets with low TDI can still achieve high performance when using BERT-based models, as these embeddings are less affected by topic distribution. On the other hand, frequency-based embeddings tend to perform worse on datasets with lower TDI because they are more prone to encountering out-of-vocabulary (OOV) words—terms that were not present in the training subset. This issue is particularly prevalent in datasets that span a wide range of topics, where the presence of OOV words can disrupt the model's ability to accurately interpret and classify text. Consequently, this leads to a degradation in overall performance, highlighting the limitations of frequency-based embeddings in handling diverse lexical elements across multifaceted thematic domains.

In the context of frequency-based text analysis frameworks, Average Inverse Document Frequency (IDF) is a crucial quality factor that highlights the predictive strength of words within a dataset. A higher average IDF indicates the presence of terms that, while rare across the entire dataset, are significantly concentrated within specific classes. These rare terms possess strong predictive capabilities due to their selective distribution, making them highly effective at distinguishing between classes. In frequency-based setups, a higher average IDF signals the presence of robust lexical predictors that enhance the model's ability to differentiate between class labels with greater precision. This metric thus provides valuable insights into the discriminative power of the dataset's features, identifying terms that are most indicative of class membership.

The correlation between dataset size and performance metrics, such as accuracy and F1-score, reveals notable dynamics in text classification models. Accuracy, which measures the overall proportion of correct predictions, shows little to no correlation with dataset size, as it does not account for how well the model handles different types of classification errors. In contrast, the F1-score, which balances precision and recall, positively correlates with dataset size. Larger datasets offer a broader range of examples, enhancing the model's ability to accurately classify instances from less prevalent classes. As dataset size increases, models become more adept at capturing class-specific patterns, leading to improvements in F1-score, which reflect better precision and recall. This is especially important in scenarios with class imbalances, where distinguishing minority classes is crucial.

6. Strengths and weaknesses

This study significantly advances the field of text classification through the introduction of new Quality Factors such as the Class Distribution Index (CDI) and the Topic Distribution Index (TDI). These innovative metrics provide a structured approach to assessing and optimizing dataset quality, ensuring that datasets are well-suited for training sophisticated machine learning models. By defining optimal QF settings, the research not only guides dataset collection but also enhances the predictive power and reliability of sentiment analysis models. These contributions are pivotal in setting new standards for dataset evaluation, thereby improving the accuracy and applicability of text classification across various domains.

Notwithstanding these insights, this study encounters several limitations that urge a cautious interpretation of its findings and highlight the need for broader research. Analyzing only six datasets may not fully capture diverse linguistic patterns, and focusing solely on machine learning models while postponing deep learning exploration could limit our current understanding. Furthermore, relying exclusively on random under-sampling for data balance might miss critical data, suggesting the need for a broader array of balancing strategies.

7. Conclusion

In text classification, the quality of a dataset greatly influences how well the resulting classifiers work. A good dataset is characterized by its balance, focus on specific themes, presence of meaningful vocabulary, and large size. These characteristics collectively facilitate the development of classifiers that are not only adept at achieving high predictive accuracy but also demonstrate commendable generalization across diverse linguistic terrains. A balanced dataset ensures fair class representation, reducing bias; thematic coherence enables the extraction of pertinent patterns; lexically potent terms, indicated by high Inverse Document Frequency (IDF) values, serve as strong predictors; and a large dataset size offers a rich linguistic diversity, enhancing the training process.

The study's limitations emphasize the need for careful interpretation and pave the way for further research. Future work should expand the range of datasets to a broader range of dialects and domains, explore deep learning technologies, and assess various data balancing techniques. These steps will enhance

our understanding of dataset optimization for text classification, pushing the field towards more advanced and effective analytical methods.

We also intend to investigate whether transformer-based models, such as AraBERT and CAMeLBERT, exhibit reduced sensitivity to data imbalance and quality factors like CDI and TDI. Their contextual encoding capacity may help overcome challenges posed by distributional skew and limited vocabulary diversity in Arabic sentiment corpora.

References

1. Touri Othmane, Sanaa El Filali, and El Habib Benlahmar. "Investigating the Effectiveness of Multilingual NLP Models for Sentiment Analysis." In: World Academy of Science, Engineering and Technology (WASET) Conference. Abstract. WASET, Oct. 2023. URL: https://publications.waset.org/abstracts/166284/investigating-the-effectiveness-of-multilingual-nlp-models-for-sentiment-analysis.
2. Mariame Tarsi, Samira Douzi, and Abdelaziz Marzak. "Predicting stock price using LSTM and Social Media dataset." In: *2023 3rd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*. 2023 3rd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET). May 2023, pp. 1–4. DOI: 10.1109/IRASET57153.2023.10152930. URL: https://ieeexplore.ieee.org/abstract/document/10152930 (visited on 03/21/2024).
3. Girma Yohannis Bade, Olga Kolesnikova, and Jose Luis Oropeza. "Evaluating the Quality of Data: Case of Sarcasm Dataset." In: (2024).
4. Ayoub Jannani, Nawal Sael, and Faouzia Benabbou. "Machine learning for the analysis of quality of life using the World Happiness Index and Human Development Indicators." In: *Mathematical Modeling and Computing* (May 2023), p. 534. DOI: 10.23939/mmc2023.02.534. URL: https://science.lpnu.ua/mmc/all-volumes-and-issues/volume-10-number-2-2023/machine-learning-analysis-quality-life-using (visited on 03/22/2024).
5. Chaimae Zaoui, Faouzia Benabbou, and Ettaoufik Abdelaziz. "Edge-Fog-Cloud Data Analysis for eHealth-IoT." In: *International Journal of Online and Biomedical Engineering (iJOE)* 19 (June 2023), pp. 184–199. DOI: 10.3991/ijoe.v19i07.38903.
6. Kamel Jafar and Panov Alexander. "Sentiment Analysis of Arabic Tweets Using SVM Classifier with POS Tagging Features." In: *International Journal of Open Information Technologies* 11.6 (May 2023). Number: 6, pp. 29–37. ISSN: 2307-8162. URL: http://injoit.ru/index.php/j1/article/view/1523 (visited on 06/16/2023).
7. Pinar Savci and Bihter Das. "Prediction of the customers' interests using sentiment analysis in e-commerce data for comparison of Arabic, English, and Turkish languages." In: *Journal of King Saud University - Computer and Information Sciences* 35.3 (Mar. 2023), pp. 227–237. ISSN: 1319-1578. DOI: 10.1016/j.jksuci.2023.02.017. URL: https://www.sciencedirect.com/science/article/pii/S131915782300054X (visited on 06/18/2023).
8. Mohamed Aly and Amir Atiya. "LABR: A Large Scale Arabic Book Reviews Dataset." In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. ACL 2013. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 494–498. URL: https://aclanthology.org/P13-2088 (visited on 06/18/2023).
9. Sarah N. Alyami and Sunday O. Olatunji. "Application of Support Vector Machine for Arabic Sentiment Classification Using Twitter-Based Dataset." In: *J. Info. Know. Mgmt.* 19.1 (Mar. 2020), p. 2040018. ISSN: 1793-6926. DOI: 10.1142/S0219649220400183. (Visited on 06/19/2023).
10. Hichem Rahab, Hichem Haouassi, and Abdelkader Laouid. "Rule-Based Arabic Sentiment Analysis using Binary Equilibrium Optimization Algorithm." In: *Arabian Journal for Science and Engineering* 48.2 (Feb. 2023), pp. 2359–2374. ISSN: 2191-4281. DOI: 10.1007/s13369-022-07198-2. (Visited on 06/23/2023).
11. Mohammed Rushdi-Saleh et al. "OCA: Opinion corpus for Arabic." In: *Journal of the American Society for Information Science* 62.10 (Oct. 2011), pp. 2045–2054. ISSN: 1532-2890. DOI: 10.1002/asi.21598. (Visited on 06/23/2023).

12. Hanane Elfaik and El Habib Nfaoui. "Deep Contextualized Embeddings for Sentiment Analysis of Arabic Book's Reviews." In: *Procedia Computer Science*. 4th International Conference on Innovative Data Communication Technology and Application 215 (Jan. 2022), pp. 973–982. ISSN: 1877-0509. DOI: 10.1016/j.procs.2022.12.100. URL: https://www.sciencedirect.com/science/article/pii/S1877050922021718 (visited on 06/18/2023).

13. Basma Alharbi et al. *ASAD: A Twitter-based Benchmark Arabic Sentiment Analysis Dataset*. arXiv.org. Nov. 2020. DOI: 10.48550/arxiv.2011.00578. arXiv: 2011.00578v3.

14. Soukaina Mihi et al. "MSTD: Moroccan Sentiment Twitter Dataset." In: *IJACSA* 11.10 (2020). ISSN: 2156-5570. DOI: 10.14569/IJACSA.2020.0111045. URL: http://thesai.org/Publications/ViewPaper?Volume=11&Issue=10&Code=IJACSA&SerialNo=45 (visited on 09/15/2022).

15. Khaled Mohammad Alomari, Hatem M. ElSherif, and Khaled Shaalan. "Arabic Tweets Sentimental Analysis Using Machine Learning." In: *Advances in Artificial Intelligence: From Theory to Practice*. International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Springer, Cham, 2017, pp. 602–610. DOI: 10.1007/978-3-319-60042-0_66. (Visited on 06/25/2023).

16. Zhiqiang Shi and Ruchit Agrawal. "A comprehensive survey of contemporary Arabic sentiment analysis: Methods, Challenges, and Future Directions." In: *arXiv preprint arXiv:2502.03827* (2025).

17. Mohamed Zouidine and Mohammed Khalil. "Large Language Models for Arabic Sentiment Analysis and Machine Translation." In: *Engineering, Technology & Applied Science Research* 15.2 (2025), pp. 20737–20742.

18. Mikhail Krasitskii et al. "Comparative Approaches to Sentiment Analysis Using Datasets in Major European and Arabic Languages." In: *arXiv preprint arXiv:2501.12540* (2025).

19. Jalel Akaichi. "Sentiment Classification: Facebook' Statuses Mining in the "Arabic Spring" Era." In: *Artificial Intelligence: Concepts, Methodologies, Tools, and Applications*. IGI Global, 2017, pp. 1858–1883. ISBN: 978-1-5225-1759-7. DOI: 10.4018/978-1-5225-1759-7.ch076. URL: https://www.igi-global.com/chapter/sentiment-classification/www.igi-global.com/chapter/sentiment-classification/173406 (visited on 06/23/2023).

20. Ashraf Elnagar, Yasmin S. Khalifa, and Anas Einea. "Hotel Arabic-Reviews Dataset Construction for Sentiment Analysis Applications." In: *Intelligent Natural Language Processing: Trends and Applications*. Springer, Cham, 2018, pp. 35–52. DOI: 10.1007/978-3-319-67056-0_3. (Visited on 06/23/2023).

21. Areeb Alowisheq et al. "MARSA: Multi-Domain Arabic Resources for Sentiment Analysis." In: *IEEE Access* 9 (2021), pp. 142718–142728. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3120746. URL: https://ieeexplore.ieee.org/document/9576756/ (visited on 09/15/2022).

22. Mahmoud Nabil, Mohamed Aly, and Amir Atiya. "ASTD: Arabic Sentiment Tweets Dataset." In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, 2015, pp. 2515–2519. DOI: 10.18653/v1/D15-1299. URL: http://aclweb.org/anthology/D15-1299 (visited on 09/15/2022).

23. Donia Gamal et al. "Implementation of Machine Learning Algorithms in Arabic Sentiment Analysis Using N-Gram Features." In: *Procedia Computer Science*. Proceedings of the 9th International Conference of Information and Communication Technology [ICICT-2019] Nanning, Guangxi, China January 11-13, 2019 154 (Jan. 2019), pp. 332–340. ISSN: 1877-0509. DOI: 10.1016/j.procs.2019.06.048. URL: https://www.sciencedirect.com/science/article/pii/S1877050919308178 (visited on 06/15/2023).

24. Kheireddine Abainia, Siham Ouamour, and Halim Sayoud. "A novel robust Arabic light stemmer." In: *Journal of Experimental & Theoretical Artificial Intelligence* 29.3 (May 2017), pp. 557–573. ISSN: 1362-3079. DOI: 10.1080/0952813X.2016.1212100. (Visited on 09/15/2022).

25. AbdelRahim. A. Elmadany, Hamdy Mubarak, and Walid Magdy. "ArSAS: An Arabic Speech-Act and Sentiment Corpus of Tweets." In: (), p. 6.

26. Mohcine Maghfour and Abdeljalil Elouardighi. "Standard and Dialectal Arabic Text Classification for Sentiment Analysis." In: *Model and Data Engineering*. Ed. by El Hassan Abdelwahed et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 282–291. ISBN: 978-3-030-00856-7. DOI: 10.1007/978-3-030-00856-7_18.

27. Abdeljalil Elouardighi, Mohcine Maghfour, and Hafdalla Hammia. "Collecting and Processing Arabic Facebook Comments for Sentiment Analysis." In: *Model and Data Engineering*. Ed. by Yassine Ouhammou et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2017, pp. 262–274. ISBN: 978-3-319-66854-3. DOI: 10.1007/978-3-319-66854-3_20.

28. Yassir Matrane, Faouzia Benabbou, and Zineb Ellaky. "Enhancing Moroccan Dialect Sentiment Analysis through Optimized Preprocessing and transfer learning Techniques." In: *IEEE Access* (2024).

29. Leonie Weissweiler and Alexander Fraser. "Developing a Stemmer for German Based on a Comparative Analysis of Publicly Available Stemmers." In: *Language Technologies for the Challenges of the Digital Age*. Ed. by Georg Rehm and Thierry Declerck. Vol. 10713. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 81–94. ISBN: 978-3-319-73705-8. DOI: 10.1007/978-3-319-73706-5_8. (Visited on 10/09/2023).

30. K. Taghva, R. Elkhoury, and J. Coombs. "Arabic stemming without a root dictionary." In: *International Conference on Information Technology: Coding and Computing (ITCC'05) - Volume II*. International Conference on Information Technology: Coding and Computing (ITCC'05) - Volume II. Las Vegas, NV, USA: IEEE, 2005. ISBN: 978-0-7695-2315-6. DOI: 10.1109/ITCC.2005.90. URL: http://ieeexplore.ieee.org/document/1428453/ (visited on 10/09/2023).

31. Chris D. Paice. "Another stemmer." In: *SIGIR Forum* 24.3 (Nov. 1990), pp. 56–61. ISSN: 0163-5840. DOI: 10.1145/101306.101310. (Visited on 10/09/2023).

32. M. F. Porter. "An algorithm for suffix stripping." In: (). DOI: 10.1108/eb046814.

33. Yassir Matrane, Faouzia Benabbou, and Nawal Sael. "A systematic literature review of Arabic dialect sentiment analysis." In: *Journal of King Saud University - Computer and Information Sciences* 35.6 (June 2023), p. 101570. ISSN: 1319-1578. DOI: 10.1016/j.jksuci.2023.101570. URL: https://linkinghub.elsevier.com/retrieve/pii/S1319157823001246 (visited on 03/21/2024).

34. Yassir Matrane, Faouzia Benabbou, and Nawal Sael. "Sentiment analysis through word embedding using AraBERT: Moroccan dialect use case." In: *2021 International Conference on Digital Age & Technological Advances for Sustainable Development (ICDATA)*. 2021 International Conference on Digital Age & Technological Advances for Sustainable Development (ICDATA). Marrakech, Morocco: IEEE, June 2021, pp. 80–87. ISBN: 978-1-66542-901-6. DOI: 10.1109/ICDATA52997.2021.00024. URL: https://ieeexplore.ieee.org/document/9588044/ (visited on 09/15/2022).