DCSA: Darija Comments Sentiment Analysis

National School of Applied Science Al Hoceima
Data Engineering Field
Author: MOTASSIM Hamza
Supervision: KHAMJANE Aziz

ABSTRACT

Sentiment analysis, a subfield of natural language processing (NLP), has gained significant attention in recent years due to its wide range of applications, including social media monitoring, customer feedback analysis, and market research. This paper focuses on sentiment analysis of Darija, the Moroccan Arabic dialect, leveraging machine learning techniques. The objective is to develop a robust sentiment analysis model capable of understanding and classifying the sentiment expressed in Darija text.

INTRODUCTION

While extensive research has been conducted in sentiment analysis for major languages, there is a noticeable gap when it comes to non-standardized dialects such as Darija, the Moroccan Arabic dialect. Darija, spoken widely in Morocco, is a unique linguistic context that poses challenges for sentiment analysis due to its informal nature, dialectal variations, and the presence of code-switching with other languages. This research addresses these challenges by developing a sentiment analysis model tailored to the specific linguistic nuances of Darija. The study employs a diverse dataset collected from various online sources, ensuring a representative sample of Darija text. Sentiment annotation is performed to create a labeled dataset for supervised learning. In this paper, we explore and compare the performance of different machine learning algorithms for Darija sentiment analysis, aiming to identify the most effective approach. This research contributes to the broader field of sentiment analysis by extending the applicability of machine learning techniques to non-standardized dialects. The findings are expected to enhance our understanding of sentiment expression in Darija and provide valuable insights for applications in social media monitoring, customer feedback analysis, and beyond.

RELATED WORKS

While sentiment analysis has been extensively studied for major languages, research specific to non-standardized dialects is relatively limited. Existing literature often focuses on sentiment analysis for languages with standardized forms, neglecting the unique challenges posed by dialectal variations. Some notable exceptions are the work by ASTD*, which addressed an Arabic Sentiment Tweets Dataset, and OMCD* that offers an Offensive Moroccan Comments Dataset*. Another one is Sentiment Analysis Challenges of Informal Arabic Language paper. Recent advancements in machine learning techniques have shown promise in sentiment analysis tasks. Transfer learning models, such as BERT* (Bidirectional Encoder Representations from Transformers), have demonstrated state-of-the-art performance in sentiment analysis for various languages. However, their applicability to Darija remains limited too (araBert, dziriBert). This research builds upon the existing literature by specifically targeting sentiment analysis in Darija. We aim to contribute to the growing body of knowledge in sentiment analysis for non-standardized dialects, paving the way for improved understanding and applications in real-world scenarios.

DATASET DESCRIPTION

1. Data Collection

The sentiment analysis project on Darija utilizes a diverse dataset compiled from various sources to capture the unique linguistic characteristics and sentiment expressions in Moroccan Arabic. The dataset comprises five existing datasets gathered from different sources, with a focus on Darija sentiment analysis. However, three primary datasets were selected for further analysis and model training:

Moroccan Arabic Sentiment Analysis Corpus: 2k tweets [review,rating].csv:*
This dataset, originally in .arff format but modified to .csv, consists of 2,000 tweets written in Moroccan Arabic. Each tweet is associated with a review and a rating, providing valuable labeled data for sentiment analysis.
Arabic Sentiment Tweets Dataset 10k tweets [review<tab>rating].txt:*
Comprising 10,000 tweets in Arabic, this dataset includes reviews and ratings. Although not specific to Darija, its inclusion adds diversity to the dataset and enriches the training process with a broader set of Arabic sentiment expressions.
Offensive Moroccan Comments Dataset - Part1 6.4k text [,comment,off].csv and - Part2 1.6k text [,comment,off].csv:*
Split into two parts, this dataset contains a total of 8,000 comments labeled for offensive content. The inclusion of offensive comments addresses the project's focus on detecting negative and offensive sentiment in Darija text.

Datasets such as "Arabic-100k-reviews*" and "The Holy Quran" were excluded for specific reasons, such as containing more general Arabic content, which may not align with the project's objective of sentiment analysis in Moroccan Arabic. Notably, rows associated with religious text, particularly excerpts from the Quran*, were excluded from the dataset. This decision was made not only due to the potential sensitivities surrounding religious content but primarily because altering or analyzing sentiments within religious verses goes beyond the project's scope. In religious contexts, it is imperative to preserve the sanctity and unchanged nature of the Quranic verses, precluding them from sentiment analysis.

2. Data Preprocessing

Data preprocessing plays a crucial role in the success of sentiment analysis models. In this section, we outline the steps taken to clean and prepare the dataset for effective model training.

- Basic Cleaning

The initial phase of data preprocessing involves fundamental cleaning operations. The `basic_cleaning` function was employed to ensure the integrity and quality of the dataset. Specifically, the following steps were performed:

- Rows with missing values in the text data column were removed.
- Rows with empty strings in the text data column were excluded.
- Duplicate rows were eliminated to prevent biases in model training.

- Class Removal

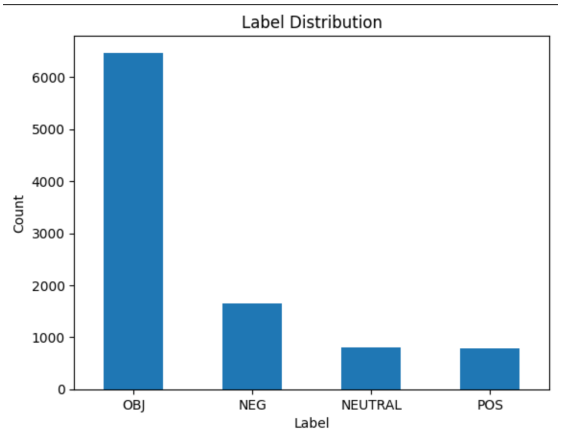
The dataset included instances that were deemed irrelevant to the sentiment analysis task. The `remove_class` function was applied to filter out rows containing specific classes. For instance, rows associated with religious text were removed to maintain respect and adhere to religious principles, instances containing religious text were omitted from the dataset. This ensures the sentiment analysis model focuses exclusively on sentiments

expressed in everyday Darija communication, aligning with the project's objectives.

- Feature Removal

Certain columns were deemed non-contributory to the sentiment analysis task and were removed from the dataset using the `remove_column` function. This step aimed to streamline the dataset and eliminate unnecessary information that could potentially introduce noise during model training.

- Balancing the Dataset



Imbalanced datasets can negatively impact the performance of machine learning models, particularly in tasks where class distribution is skewed. The `balance_dataframe` function was employed to address this issue. This function allows for either oversampling the minority class or undersampling the majority class, ensuring a balanced representation of sentiment classes in the dataset. By opting for undersampling, we aimed to address the class imbalance while optimizing computational resources and model training efficiency.

The resulting dataset, after undergoing these preprocessing steps, consisted of 12,248 rows with two classes: negative/offensive labeled “1” comments and positive/neutral labeled “0” comments. The systematic and comprehensive preprocessing approach lays the groundwork for the subsequent development and evaluation of sentiment analysis models tailored to the nuances of Darija language expressions.

- Advanced Cleaning

In addition to the basic cleaning procedures, advanced cleaning steps were applied to enhance the quality and interpretability of the text data. These advanced cleaning functions address specific linguistic and contextual nuances present in Darija, the Moroccan Arabic dialect.

- Emoji Translation

The `create_emoji_mapping` and `translate_emoji` functions were employed to handle emojis in the text data. Emojis were mapped to their corresponding textual labels using a predefined mapping dictionary. The `translate_emoji` function then replaced emojis in the text with their respective labels. This step ensures that emojis contribute to the sentiment analysis task by being interpreted as meaningful textual elements.*

	emoji	text
0		جاك فانوس
1		شجرة عيد الميلاد
2		العاب ناريه
3		الماسة
4		مفرقعة نارية

- Hashtag Normalization

The `normalize_hashtags` function was implemented to process hashtags in the text data. This involved removing the '#' symbol and replacing underscores with spaces. By normalizing hashtags in this manner, the words within the hashtags are preserved, contributing to a more accurate representation of the language.

- Text Cleaning Pipeline

The `clean_text` function orchestrates a comprehensive pipeline of text processing steps, including:

➔ Removing URLs to eliminate extraneous web links.
➔ Translating emojis using the previously created emoji mapping.
➔ Normalizing digits by converting Eastern/Persian numbers to Western numbers.

➔ Processing and cleaning hashtags.
➔ Removing user mentions (@user) to prevent potential biases.
➔ Retaining only Arabic characters.
➔ Stripping Arabic diacritics, tashkeel, and tatweel for consistency and readability.
➔ Removing repeated characters for a cleaner representation.
➔ Eliminating specific Arabic punctuation marks.
➔ Stripping extra whitespaces for improved text readability.

- English to Arabic Letter Translation

The `derrej` function facilitates the translation of English letters into their corresponding Arabic counterparts. This step is crucial for handling code-switching or instances where English letters are used to represent Arabic words.

- Text Language Detection

The `detect_text_lang` function assesses whether the input text is predominantly in Arabic, English, or a mix of both. This detection is based on the percentage of Arabic and English letters in the text. This information is valuable for subsequent language-specific processing.

- Stop Words Removal

A custom `stop_words_removal` function is used to remove stop words from tokenized sentences. Stop words were compiled from various sources, including a manually curated list, Doda API for thematic stop words, and common Arabic stop words. Thematic stop words were obtained from the Doda API for specific categories such as numbers, food, clothes, colors, pronouns, and adverbs. This enriches the stop word list with contextually relevant terms that may not be covered by generic stop word lists. The integration of these advanced cleaning steps ensures that the dataset is meticulously processed, capturing the intricacies of sentiment expressions in Darija while addressing specific linguistic characteristics and contextual nuances. This refined dataset is poised for further analysis and model training in the subsequent stages of the sentiment analysis project.*

- Dataset Stemming

In the context of natural language processing (NLP), stemming is a technique employed to reduce words to their base or root form, thereby simplifying the textual data for analysis. Stemming is particularly useful in sentiment analysis tasks to enhance computational efficiency and reduce the dimensionality of the dataset. The stemming process applied to the cleaned text in this project involved the utilization of the FarasaStemmer.

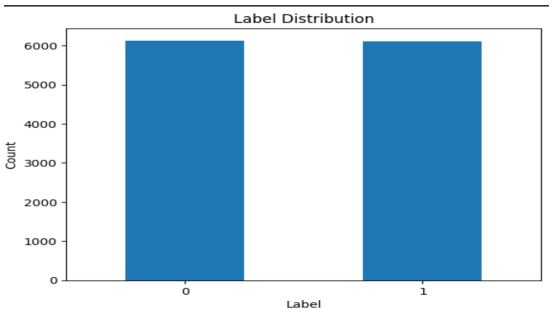
The stemming process was implemented within a loop that iterates through the 'cleaned_text' column of the DataFrame. Each cleaned text was tokenized using the Araby library, and the FarasaStemmer was applied to each token. The resulting tokens were then filtered to exclude stop words, and the processed tokens were joined back into a coherent sentence. This sequence of operations was performed for a subset of the dataset (1000 rows) for testing purposes due to the time-intensive nature of stemming(~33H).

- Time-Saving Considerations

Recognizing the substantial time requirements of stemming, a strategic decision was made to process only a subset of the dataset. Every tenth row was processed, resulting in a balance between computational efficiency and linguistic depth.

3. Dataset Summary

The DataFrame was then saved to a CSV file ('cleaned_data.csv') for future reference. The final dataset comprises 12,243 rows after dropping any rows with missing values in the 'cleaned_text' column, providing a concise yet linguistically informed representation of the sentiment expressions in Darija. The inclusion of the 'stemmed_text' column enhances the dataset's suitability for subsequent stages of sentiment analysis model training and evaluation.



The dataset lies in its inherent diversity, capturing a rich tapestry of linguistic variations and thematic nuances inherent in Darija, the Moroccan Arabic dialect. The dataset incorporates both Arabic-infused Darija and pure Darija expressions, and also encompasses a broad spectrum of topics and contexts. This diversity lends a robust and comprehensive foundation to the sentiment analysis project. It accommodates variations of Darija that are influenced by Arabic, highlighting the natural code-switching and language blending often observed in everyday communication. Simultaneously, it includes instances of pure Darija, providing a closer look at the authentic colloquial expressions as spoken by individuals in Casablanca and its nearshore regions.

index ▲	text	class	cleaned_text	stemmed_text
0	ملوال حياتي لم المس اي تغير حتى قدمت هذه الحكومة قل ثقف بجانها بصوتنا	0	ملوال حياتي لم المس اي تغير حتى قدمت هذه الحكومة قل ثقف بجانها بصوتنا	ملوال حياة لم مس تغير قدم حكومة قل جان صوت
1	منتوج رائع وثمن مناسب ....جميل	0	منتوج رائع وثمن مناسب جميل	منتوج رائع ثمن مناسب جميل
2	كلنا ابن كيران لمتاقق معايا يدير جيم	0	كلنا ابن كيران لمتاقق معايا يدير جيم	كلنا ابن كير متاقق معايا أدار
3	وفقك الله لولاية اخرى حقاش مكيناش محسن منك	0	وفقك اله لولاية اخرى حقاش مكيناش محسن منك	وفق اله ولاية أخرى حقاش مكيناش محسن
4	لأنه و بكل بساطة رئيس الحكومة يعتني بمعاق داخل بيته جعله اله في ميزان حسناته	0	لأنه و بكل بساطة رئيس الحكومة يعتني بمعاق داخل بيته جعله اله في ميزان حسناته	بساطة رئيس حكومة اعتنى معاق داخل بيت اله ميزان حسنة
5	إين الريف العظيم تحية لأبناء سلوان	0	إين الريف العظيم تحية لأبناء سلوان	إين ريف عظيم تحية ابن سلوان
6	مشاء الله الفاتحة بتقليد الشيخ ياسر الدوسري و سورة القيامة بتقليد الشيخ إدريس ابكر	0	مشاء اله الفاتحة بتقليد الشيخ ياسر الدوسري و سورة القيامة بتقليد الشيخ إدريس ابكر	مشاء اله فاتحة تقليد شيخ ياسر دوسري سورة قيامة تقليد شيخ إدريس ابكر
7	تحياتي لك ادمين على هاد المجهود لي كادير باش تقربنا من اللاعبيين ديالنا	0	تحياتي لك ادمين على هاد المجهود لي كادير باش تقربنا من اللاعبيين ديالنا	تحية ادم مجهود ادير باش تقرب اعيب ديال
8	اشرف يرفع لقب مغربي عالمي رسالة واضحة بانه قد يرفع لقب روسيا	0	اشرف يرفع لقب مغربي عالمي رسالة واضحة بانه قد يرفع لقب روسيا	أشرف رفع لقب مغربي عالمي رسالة واضح رفع لقب روسيا
9	مبروك معرفتش لاش مغاربة كيهرسو ويطيحو بقيمة ولاد بلادهم ويعطو قيمة للبراني	0	مبروك معرفتش لاش مغاربة كيهرسو ويطيحو بقيمة ولاد بلادهم ويعطو قيمة لبراني	مبروك معرفتش لاش مغربي كيهرسو يطيحو قيمة ولاد بلد يعطو قيمة براني
10	المقال لا يعكس الواقع على الإطلاق	1	المقال لا يعكس الواقع على الإطلاق	مقال لا عكس واقع إطلاق
11	هذا شخص مريض شانه شان الدواعش المرضى والمهووسين بعلاقتهم مع الله هذه الفئات المريضة	1	هذا شخص مريض شانه شان الدواعش المرضى والمهووسين بعلاقتهم مع الله هذه الفئات المريضة	شخص مريض شان شان دواعش مريض مهوس علاقة اله فئة مريض
12	يا لهول!!! حتى الأمس القريب كنت أحسب أن الحمير لا علاقة لها بالفلسفة و الفكر الحر أما و الحالة هاته فسائلت الرحال إلى بقعة من كوكبنا لا يوجد بها مثل هاته الأنواع الهجينة من الحمير	1	يا لهول حتى الأمس القريب كنت أحسب أن الحمير لا علاقة لها بالفلسفة و الفكر الحر أما و الحالة هاته فسائلت الرحال إلى بقعة من كوكبنا لا يوجد بها مثل هاته الأنواع الهجينة من الحمير	هول حمار لا علاقة فلسفة فكر حر حالة أتد رحل بقعة كوكب لا نوع هجين حمار
13	مناورات روسيا في المنطقة العربية كلها بضوء أخضر أمريكي إسرائيلي من أجل إنهاك دول المنطقة	1	مناورات روسيا في المنطقة العربية كلها بضوء أخضر أمريكي إسرائيلي من أجل إنهاك دول المنطقة	مناورة روسيا منطقة عربي ضوء أخضر أمريكي إسرائيلي إنهاك دولة منطقة
14	مسرلحة لم نبق نعرف من ضند من في هذه الفوضى الهلاكة التي بدات بالربيع العبري	1	مسرلحة لم نبق نعرف من ضند من في هذه الفوضى الهلاكة التي بدات بالربيع العبري	لم نبق عرف هد فوضى هلاك بدا ربيع عبري
15	سير ا خويا الله يجيب من يستارك	0	سير ا خويا اله يجيب من يستارك	سير خوي اله أجاب ستر
16	اشنو المشكيل ديالها	0	اشنو المشكيل ديالها	اشنو مشكيل ديال
17	أول كومنت	0	أول كومنت	كومن
18	العبادي يلغي مناصب السنة.	0	العبادي يلغي مناصب السنة	عبادي لغى منصب

4. MODELS

The sentiment analysis task involves the exploration and evaluation of various machine learning, deep learning models and LLM to identify the most effective approach for the specific context of Darija sentiment analysis. Different models and architectures are considered, each with its set of hyperparameters. The goal is to select the model that demonstrates optimal performance in terms of accuracy and classification metrics.

- Machine Learning Models

In the initial phase, a diverse set of traditional machine learning models, including Multinomial Naive Bayes, Logistic Regression, XGBoost, and Support Vector Machine (SVM), are explored for sentiment analysis. These models employ the TF-IDF vectorization technique to convert text data into numerical features, enhancing their efficacy. Hyperparameter tuning is conducted using GridSearchCV to fine-tune model parameters. Naive Bayes considers n-gram range, maximum document frequency, and alpha smoothing. Logistic Regression focuses on n-gram range, maximum document frequency, regularization strength (C), and penalty type. XGBoost is optimized for n-gram range, maximum document frequency, number of estimators, maximum depth, and learning rate. SVM, configured with TF-IDF vectorization, undergoes fine-tuning with parameters like n-gram range, maximum document frequency, regularization (C), and kernel choice (linear or radial basis function - RBF). The

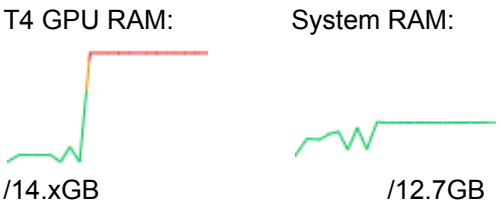
best-performing models, identified through cross-validation on the training set, are further assessed on validation and test sets, considering accuracy scores and detailed classification reports.

- Deep Learning Models

To explore the potential of deep learning in sentiment analysis, neural network architectures are considered. A range of hyperparameter sets is experimented with, encompassing variations in embedding dimensions, dropout rates, convolutional filters (CNN), recurrent units (SimpleRNN and LSTM), and architectural choices. The neural network models are trained on tokenized and padded sequences of the text data. The training process involves multiple epochs, and the models are evaluated on the test set to gauge their final performance. Classification reports and accuracy scores provide a comprehensive understanding of how well these models generalize to unseen data. The diverse set of models, spanning traditional machine learning to deep learning architectures, enables a thorough exploration of the most suitable approach for Darija sentiment analysis. The subsequent sections present the results and comparative analyses, shedding light on the strengths and limitations of each model in the context of the project's objectives.

- LLM's

In the Language Model (LLM) section, a pre-trained BERT model is utilized for sentiment analysis. The BERT model is fine-tuned on the specific sentiment analysis task using TensorFlow and the Transformers library. The sentiment analysis dataset is split into training, testing, and validation sets, and then converted into a format suitable for BERT tokenization. The dataset is processed using the BERT tokenizer, and the resulting tokenized data is formatted into TensorFlow datasets. The input features include 'input_ids', 'attention_mask', 'token_type_ids', and the target labels ('class'). The data is organized into batches, and a custom function 'order' is applied to structure the input features and labels for the BERT model. A custom BERTForClassification model is defined using TensorFlow and integrated with the pre-trained BERT model. The model is compiled with an Adam optimizer, sparse categorical cross-entropy loss function, and accuracy as the evaluation metric. The training process is carried out for a specified number of epochs.
Regarding computing limitations, it's essential to note that training deep learning models, especially large pre-trained language models like BERT, can be computationally intensive and require substantial resources. Fine-tuning BERT on a substantial dataset might demand significant processing power and memory. Depending on the available hardware (e.g., GPU or TPU), the training time and feasibility may vary. It's common to face limitations, particularly on personal machines or less powerful computing environments, where training large models might be impractical due to time constraints or hardware restrictions.



- Experiments and Results:

MODEL	BEST HYPERPARAMETERS	VALIDATION ACC	FINAL TEST ACC
MultinomialNB	{'alpha': 0.5, 'fit_prior': False, 'max_df': 0.5, 'ngram_range': (1, 2)}	0.743	**0.762**
LogisticReg	{'C': 2.0, 'penalty': 'l2', 'max_df': 0.5, 'ngram_range': (1, 1)}	0.742	0.729

XGBoost	{'learning_rate': 0.2, 'max_depth': 9, 'n_estimators': 200, 'max_df': 0.5, 'ngram_range': (1, 1)}	0.682	0.673
SVM	{'C': 1, 'kernel': 'rbf', 'max_df': 0.5, 'ngram_range': (1, 2)}	0.722	0.733
CNN	{'embedding_dim': 50, 'dropout_rate': 0.2, 'conv_filters': 32, 'kernel_size': 3, 'architecture': 'CNN'}	-	0.660
RNN	{'embedding_dim': 100, 'dropout_rate': 0.3, 'recurrent_units': 16, 'architecture': 'SimpleRNN'}	-	0.678
LSTM	{'embedding_dim': 50, 'dropout_rate': 0.3, 'recurrent_units': 16, 'architecture': 'LSTM'}	-	0.675
distilBert	{'Batch Size': 8, 'Epochs': 1, 'Learning Rate': 1e-2}	-	**0.928**

5. ERROR ANALYSIS:

- Data Errors:

Darija presents several challenges and considerations in natural language processing (NLP) tasks. Here are some common constraints and challenges associated :

Dialectical Variation: " نتا vs نتايا vs نتينا "

Code-Switching: " crazy man had khona hada "

Lack of Standardization: "حمام"

darija	7mam	7mmam
english	pigeons	bathroom

Named Entity Recognition (NER) Challenges:

"مشات فاطمة للمدرسة"

For error analysis, it's important to consider the types of mistakes your model makes. Common error types might include misinterpretation of code-switched phrases, difficulty in handling dialectical variations, or inaccuracies in recognizing named entities. Error analysis can provide insights into areas that need improvement and guide further model development. Additionally, collecting and annotating more diverse and representative datasets in Darija can contribute to addressing some of these challenges.

6. CONCLUSION:

In this study, we undertook a comprehensive exploration of sentiment analysis in Darija, the Moroccan Arabic dialect. The research aimed to address the unique linguistic challenges posed by Darija and contribute to the broader field of natural language processing in non-standardized and informal linguistic contexts.

Our sentiment analysis experiments revealed the efficacy of traditional machine learning models, such as Multinomial Naive Bayes, Logistic Regression, XGBoost, and Support Vector Machine, in capturing sentiment nuances in Darija text. The models, tuned using hyperparameter optimization, demonstrated promising accuracy levels in both validation and test datasets.

Furthermore, the integration of deep learning models, specifically the implementation of BERT-based models like DistilBERT, showcased significant improvements in sentiment analysis tasks. The transfer learning capabilities of these models, fine-tuned on Darija datasets, underscored their potential for handling the intricacies of informal and non-standard Arabic dialects.

However, it is crucial to acknowledge certain limitations, including the need for larger and more diverse datasets, potential biases in training data, and computational constraints in fine-tuning sophisticated models.

In conclusion, this research contributes valuable insights into sentiment analysis and named entity recognition in Darija, offering a foundation for future work in developing more robust and culturally sensitive natural language

processing models for Arabic dialects. As technology evolves, addressing linguistic diversity and informal language usage becomes imperative for the advancement of natural language processing applications in real-world, multicultural contexts.

REFERENCES

Moroccan Arabic Sentiment Analysis Corpus: https://github.com/ososs/Arabic-Sentiment-Analysis-corpus/blob/master/MSAC.arff
ASTD: https://github.com/mahmoudnabil/ASTD/blob/master/data/Tweets.txt
GOUD: https://huggingface.co/datasets/Goud/Goud-sum
MNAD: https://www.kaggle.com/datasets/jmourad100/mnad-moroccan-news-articles-dataset
QCRI: https://github.com/qcri/QADI
OMCD: https://github.com/kabilessefar/OMCD-Offensive-Moroccan-Comments-Dataset
The Holy Quran: https://www.kaggle.com/datasets/zusmani/the-holy-quran
100k books & hotels: https://www.kaggle.com/datasets/abedkhooli/arabic-100k-reviews
emoji-to-arabic-text: https://www.kaggle.com/datasets/hatemamine/emoji-to-arabic-text
Stop Words: 800
https://github.com/mohataher/arabic-stop-words/blob/master/list.txt
Doda: https://github.com/darija-open-dataset/dataset/
ArSAS: http://lrec-conf.org/workshops/lrec2018/W30/pdf/22_W30.pdf
SemEVAL: https://aclanthology.org/S17-2088.pdf
distilBert: https://huggingface.co/distilbert-base-uncased
farssapy: https://github.com/MagedSaeed/farasapy