RESEARCH PAPER

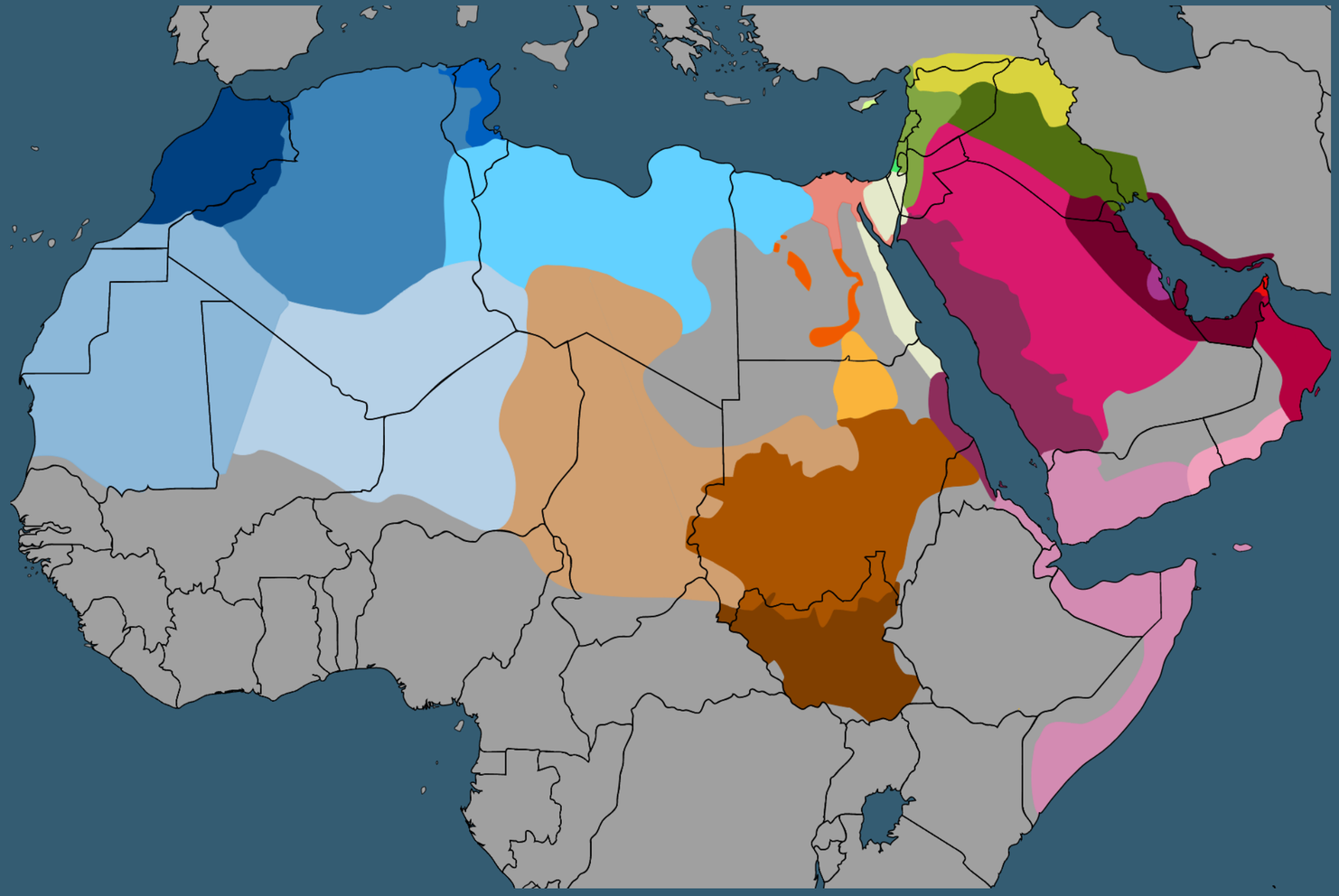
Arabic Text Dialect Recognition

Presented By:
Mohand Al-Rasheed
Khalid Albader
Abdulrahman Alshawi
Abdullah Alsuwailem
Musaad Alqubayl

Supervised by: Dr. Nasser Alsadhan

Abstract

- The Arabic Language and its variety of dialects
- Applications
- What are we building



Introduction

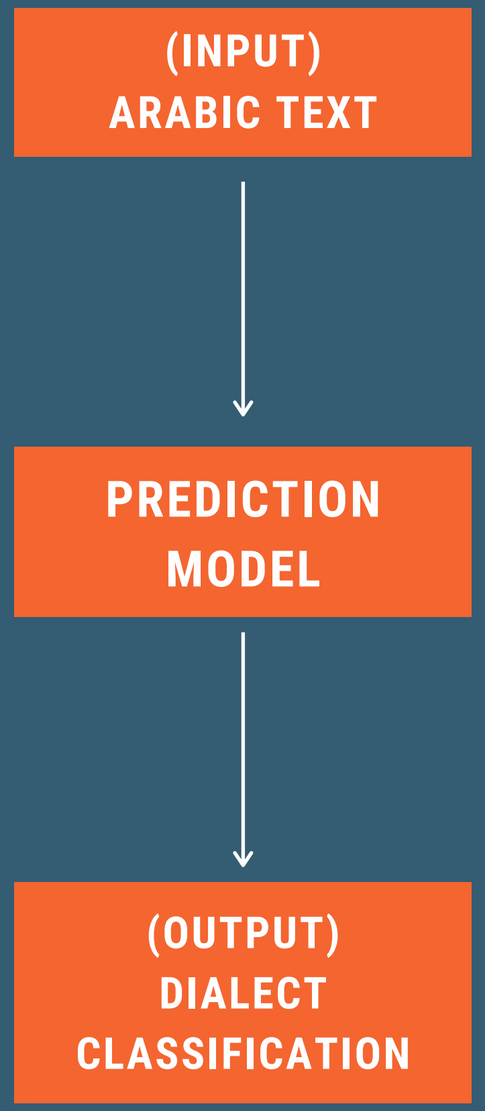
PART 1

Research Problem

WHAT WE WANT TO SOLVE

- Many Arabic speaking regions have formed dialects exclusive to their own

- Identify and predict dialect types from text.



Goals and Objectives

WHAT WE WANT TO ACHIEVE

→

Analyze and understand Arabic text to classify the dialect of any piece of Arabic text

→

Implement the most appropriate state of the art NLP model that helps in achieving the best possible accuracy

Research scope

Analyzing, preprocessing and modeling a state of the art NLP model to classify Arabic text into a set of dialects.

Review of Related Literature

PART 2

Related Literatures

SUMMARY

COTTERELL-BURCH

DATASET

They used an extended version of the Arabic Online Commentary (AOC) dataset which gathered millions of comments from three newspapers

ALGORITHM & RESULTS

They used two algorithms, SVM and Naive Bayes using unigram, bigram and trigram features. Highest results were produced by using NB UNI which achieved 87% accuracy

AREEJ ALSHUTAYRI & ERIC ATWELL

DATASET

Alshutayri used the Social Media Arabic Dialect Corpus(SMADC) dataset to classify dialects to GLF, NOR, LEV, EGY and IRQ

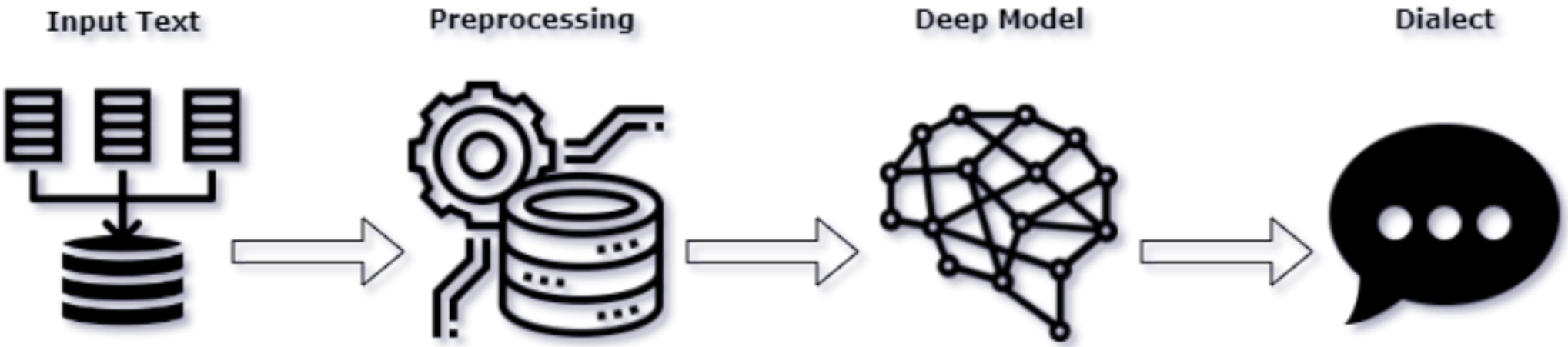
ALGORITHM & RESULTS

They used Sequential Minimal Optimization (SMO) algorithm with multinomial Naive Bayes (MNB) with different tokenizers, they achieved an accuracy of 88% using MNB-TFIDF (multinomial naive bayes with TFIDF)

Methodology

PART 3

Methodology



High-level view of inference procedure of a single input sentence

SMADC Dataset

SOCIAL MEDIA ARABIC DIALECT CORPUS

COLLECTION

SMADC's corpus is collected from three different sources, Facebook, Twitter and online newspapers

FILTERATION

Facebook and Twitter documents were filtered automatically by removing hashtags, emojies, redundant characters.

ANNOTATION

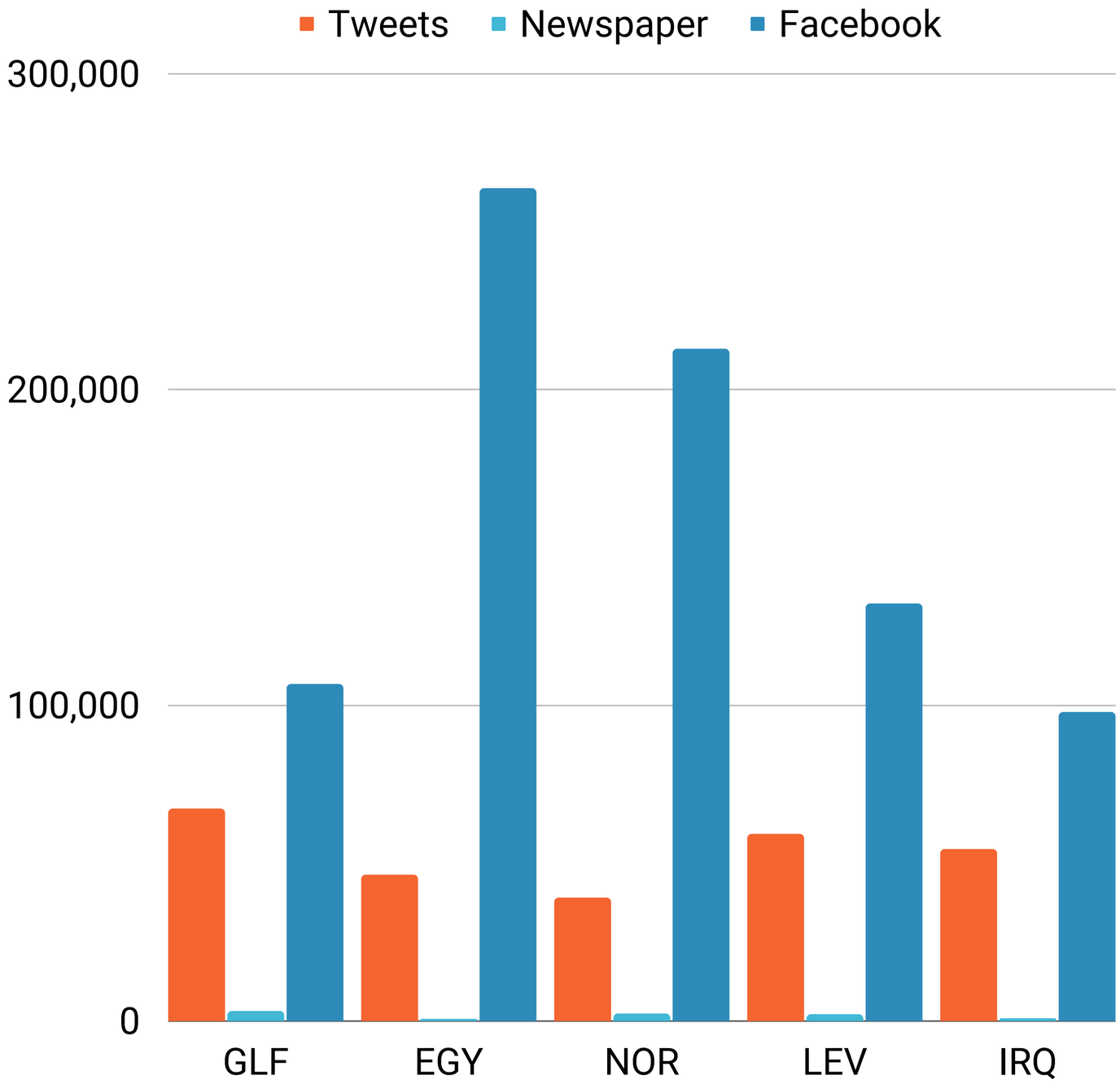
After automatically annotating the documents, the researcher has used novel manual annotation techniques to annotate a part of the dataset

FINAL VERSION

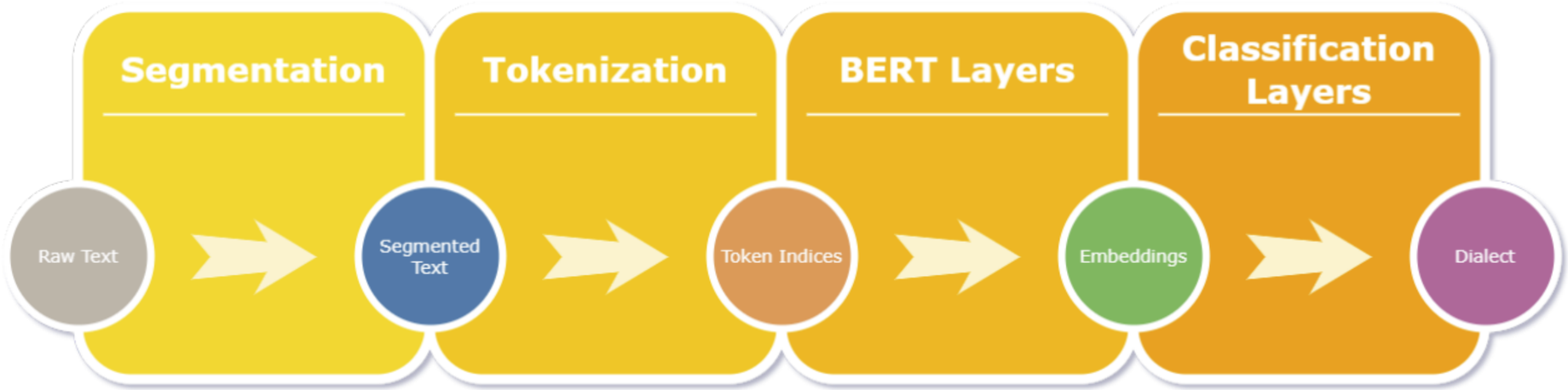
SMADC dataset contained 1,088,578 documents. which consisted of 812,849 Facebook comments, 9,440 online newspaper comments, and 266,289 Twitter tweets. And each one of them are distributed in the five labels (GLF, EGY, NOR, LEV and IRQ)

SMADC Dataset

SOCIAL MEDIA ARABIC DIALECT CORPUS



Data Preprocessing



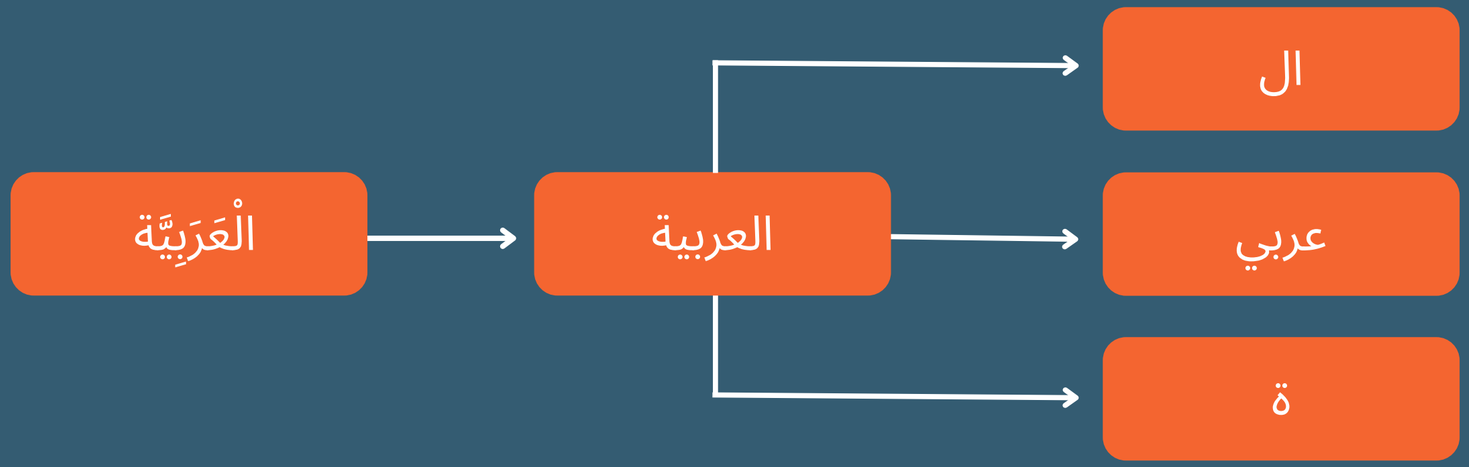
High-level view of preprocessing procedure of a single input sentence

Data Preprocessing

NORMALIZATION & SEGMENTATION → **TOKENIZATION**

Normalize Arabic diacritics such as fatha, damma, kasra and so on. segmentation works by separating the suffixes and prefixes attached to any given word

Transform each token to a number, so if one token is repeated more than once then that token is transformed to the same number



BERT

BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS

We'll be using an Arabic BERT model called AraBERT and fine-tuning it to the problem we need to solve.

We need to provide a sequence of tokens represented as numbers, after tokenizing our data we'll replace each token with a number representation, this forms a suitable input for AraBERT.

Experimental design

PART 4

Algorithms

DATA IMBALANCE

uge label imbalance in EGY and NOR labels as they dominate the dataset in comparison to IRQ and LEV labels, we intend to mitigate such imbalances by experimenting with imbalance correction techniques such as SMOTE if needed

PERFORMANCE

We'll experiment with a number of hyperparameters that impact the performance of our model. We'll also experiment with the batch size and number of epochs

OPTIMALITY

We've also introduced a warmup ratio which should accelerate our model to achieving optimality.

Results

PART 5

Results

RESEARCH FINDINGS

Model name	Accuracy
bert-large-arabertv2	0.892
bert-base-arabertv2	0.872
bert-base-arabertv02-twitter	0.826
Linear SVM	0.747
MultinomialNaiveBayes	0.865
RandomForest	0.760

results on a test set of the SMADC dataset.

Results

RESEARCH FINDINGS

	Precision	Recall	F1-Score	Training Samples	Testing Samples
EGY	0.819585	0.775428	0.796895	622108	1986
GLF	0.866476	0.838073	0.852038	180244	2532
IRQ	0.862557	0.863045	0.862801	153933	1767
LEV	0.842440	0.823602	0.832915	195631	1610
NOR	0.933374	0.966609	0.949701	256654	6319

class-specific results for the best performing model (bert-large-arabertv2)

Results

RESEARCH FINDINGS

	EGY	GLF	IRQ	LEV	NOR
EGY	0.863045	0.037351	0.028862	0.027731	0.043011
GLF	0.041793	0.775428	0.069486	0.035247	0.078046
IRQ	0.024487	0.049368	0.838073	0.034360	0.053712
LEV	0.038509	0.049689	0.045342	0.823602	0.042857
NOR	0.005697	0.010761	0.010286	0.006647	0.966609

the confusion matrix for the test set

Results

RESEARCH FINDINGS

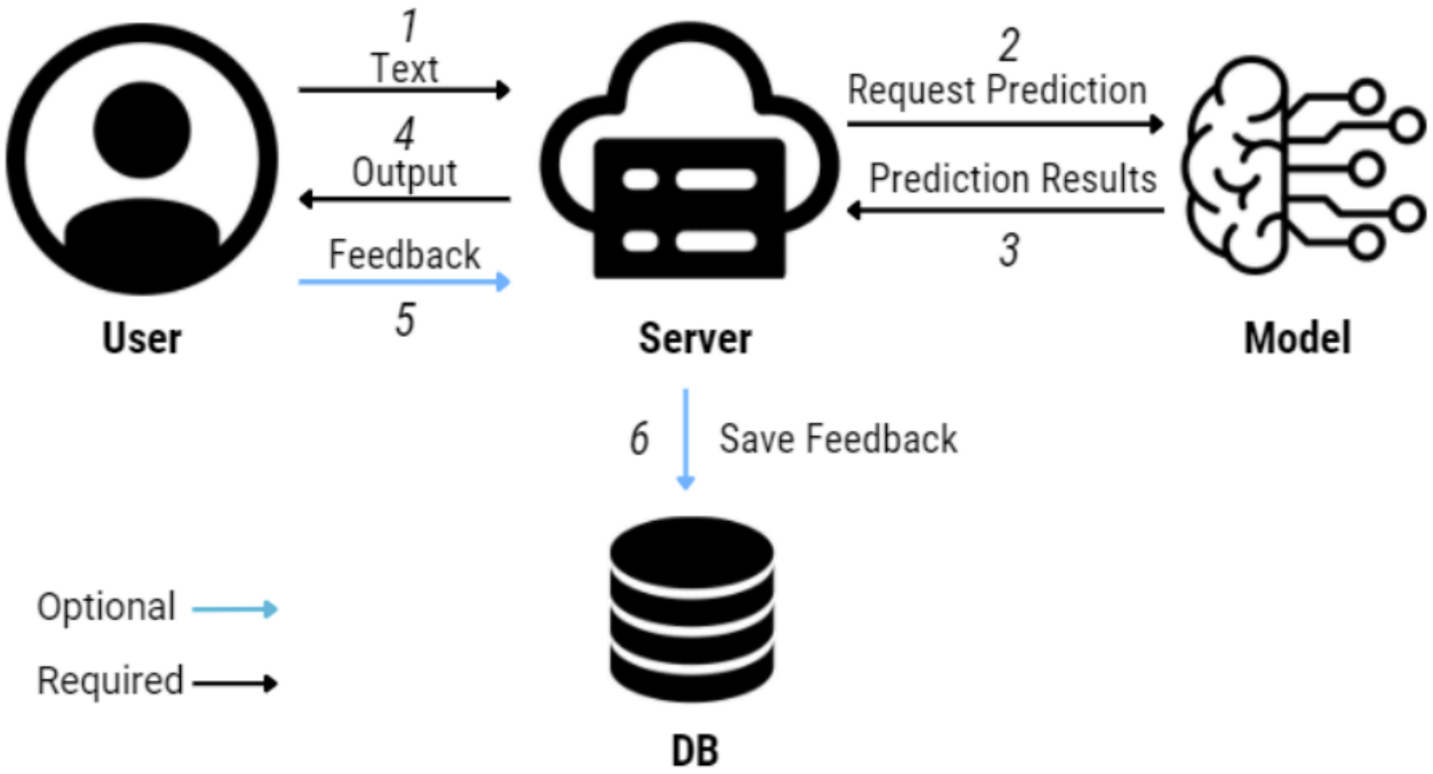
Model name	Dataset	Macro F1	Macro precision	Macro recall
MultinomialNaiveBayes	annotated_data	0.552954	0.561883	0.611622
MultinomialNaiveBayes	arabic_dialects	0.450884	0.464933	0.446934
MultinomialNaiveBayes	dart	0.737389	0.742771	0.748739

the most robust traditional models (by Macro F1)

Model name	Dataset	Macro F1	Macro precision	Macro recall
bert-large-arabertv2	annotated_data	0.608431	0.597582	0.660434
bert-large-arabertv2	arabic_dialects	0.449060	0.459210	0.440770
bert-base-arabertv2	dart	0.760816	0.766500	0.759033

the most robust BERT models (by Macro F1)

Interface



Demonstration structure

Interface

EXAMPLE

Try it:
https://arabic.hawzen.me

LIVE DEMO

Predict Region

الزمالك فريق جامد أوي

Enter text in Arabic

Is the prediction correct? YES NO

BERT BAYES

CLASSIFY