MARASTA: A Multi-dialectal Arabic Cross-domain Stance Corpus

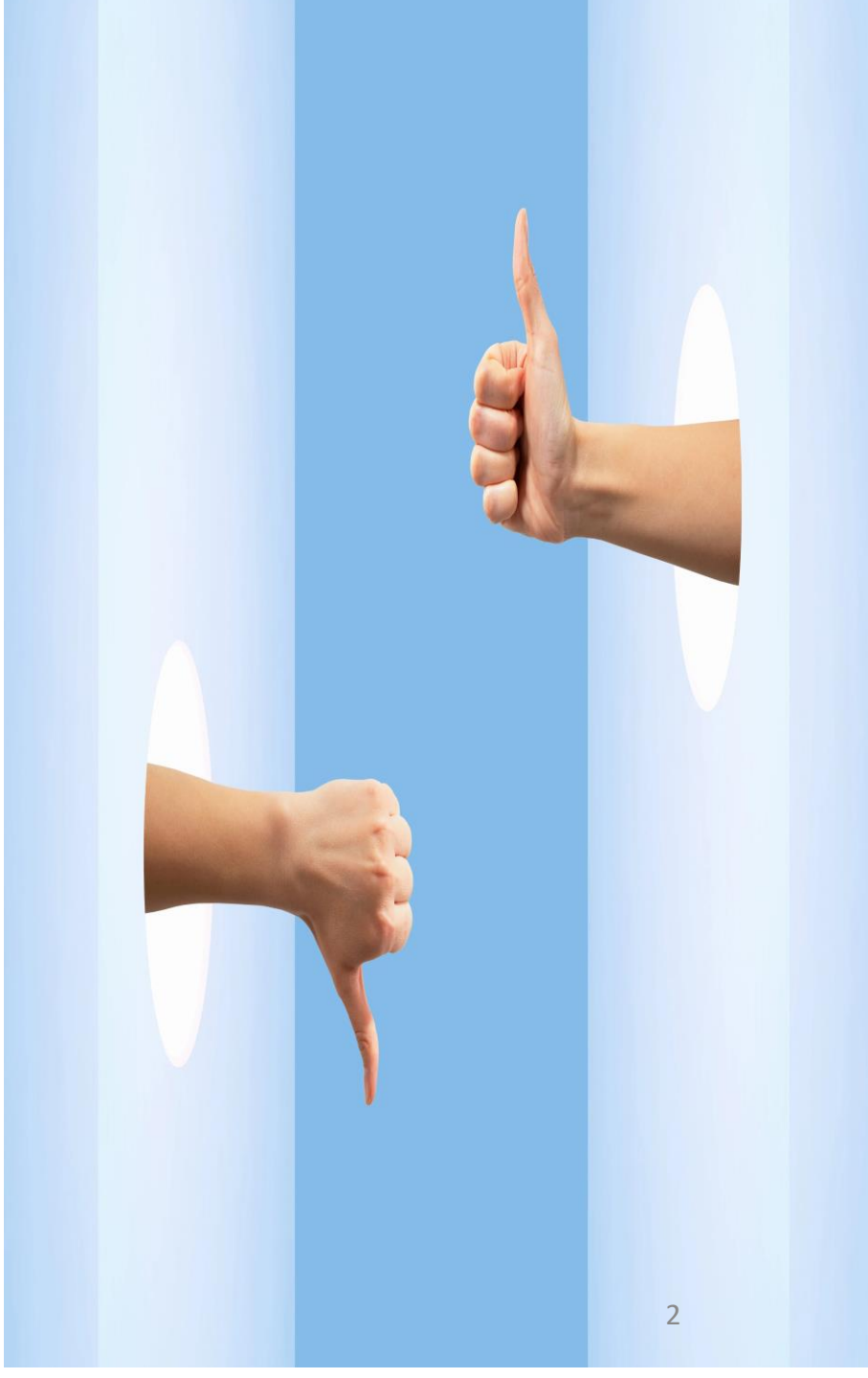
Anis Charfi , Mabrouka Bessghaier , Andria Atalla , Raghda Akasheh , Sara Al-Emadi , Wajdi Zaghouani

LREC-COLING 2024

Background – Stance detection

- The word **stance** refers to the writer expressing personal feelings, assessments, and judgments about a certain message or topic.

- Stance detection is a sub-task that branches out from sentiment analysis, which aims to determine the author's attitude towards a target, often stating whether this author is **against**, **in favor**, or **neutral** towards it.



Background – Use of stance detection

- ✓ Conducting analytical research to measure the public opinion on social, religious, and political issues.
- ✓ Veracity checking applications.
- ✓ Determining the level of controversy on social media platforms.
- ✓ Making critical decisions.
- ✓ Detecting polarization.

Arabic Dialects

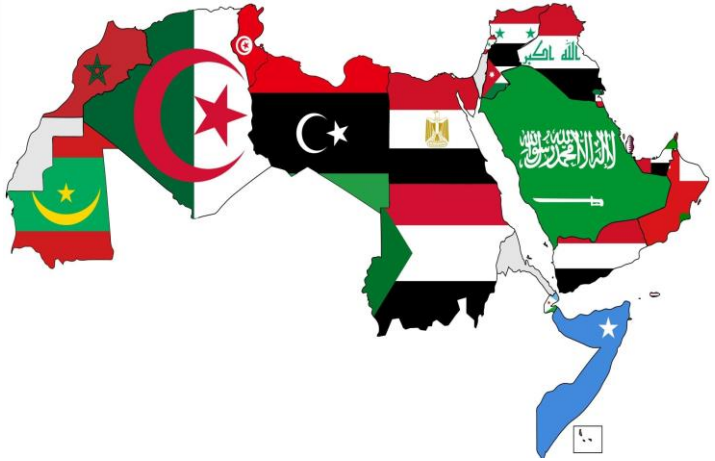
- Arabic dialects vary significantly; each dialect has unique linguistic features, syntax, and vocabulary.

Historical events

Cultural developments

Geographic isolation

Interactions with neighboring languages



Related Work

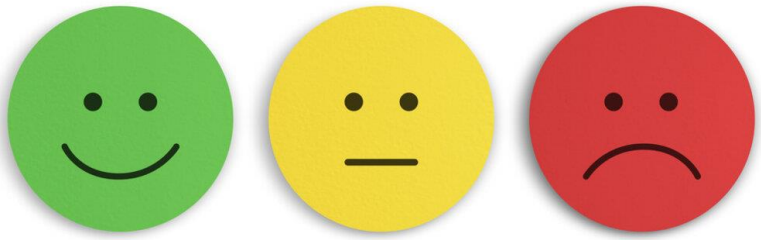
Darwish et al. (2017)	Jaziriyan et al. (2021)	(Alturayeif et al., 2022).
• Focuses on only one target	• Cover different dialects and targets • Not balanced • The corresponding dialects are not specified	• Cover different dialects and targets • Not balanced • The corresponding dialects are not specified

We propose a **balanced** dataset for Arabic stance detection that spans a **variety of topics** and covers the main Arabic **dialects**.

Proposed Work

الصندوق القطري لرعاية البحث العلمي
Qatar National Research Fund
Member of Qatar Foundation

- ✓ Cross-domain and multi-dialectal stance corpus for Arabic called **MARASTA**
- ✓ It includes over 4,500 sentences annotated concerning their stance toward a topic by at least two annotators.
- ✓ For each sentence, our annotators manually indicated if the sentence's stance favored a certain topic, against it, or neutral.



Proposed Work

- ✓ **MARASTA** covers four important dialectal regions in the Arab world: Maghreb, Egypt, the Gulf, and the Levant.
- ✓ We collected at least 1,000 sentences for each of the four Arab regions, totaling 4657 sentences.

MARASTA Corpus - Data Collection

- Our research team consisted of individuals from diverse Arab countries, each deeply understanding the controversial topics in their respective regions.

Criteria for choosing the topics

✓ A topic should be sufficiently controversial to ensure plenty of discussions, with many posts and comments expressing different stances.

✓A topic can be either current or previously pertinent (within the past 15 years).

✓ A topic can be relevant to one country or multiple countries.

MARASTA Corpus - Data Collection

Egyptian dialect — New Administrative Capital in Egypt

- العاصمة الإدارية الجديدة ; العاصمة الجديدة ; العاصمة الادارية
  (*Translation: New Administrative Capital, The new capital, The Administrative Capital*)

Maghreb — Illegal Immigration to Europe

- الحرقة ; الهجرة الغير نظامية ; الهجرة السرية : أوروبا
  (*Translation: Immigration, Irregular Immigration, Clandestine Immigration, Europe)*

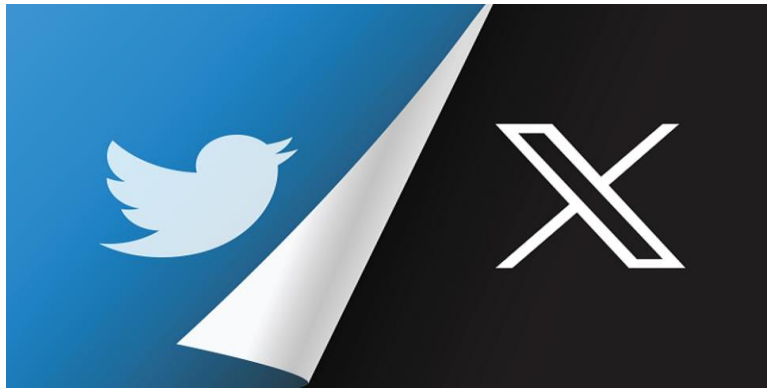
Gulf Region — Fifa World Cup 2022

- كأس العالم في قطر; تنظيم كأس العالم قطر; فيفا ٢٠٢٢ ; فيفا قطر; مونديال
  (Translation: World Cup in Qatar; The Organization of Qatar's World Cup; FIFA 2022; FIFA Qatar; Qatar's Mundial)

Levant Region — Feminism and women's Rights

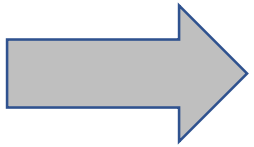
- النسويات؛ الحركة النسوية؛ الاحتجاجات النسوية؛ الذكورية
  (Translation: Feminists; The Feminist Movement; Feminist Protests; Masculinity)

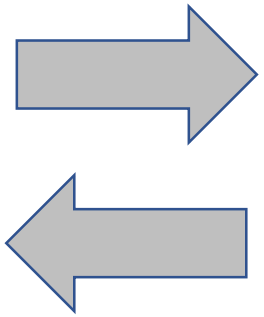
MARASTA Corpus - Data Collection

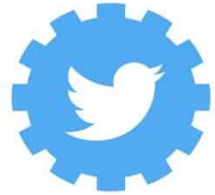


MARASTA Corpus - Data Collection

A set of keywords related to the topic

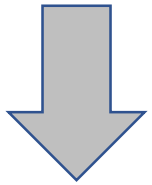






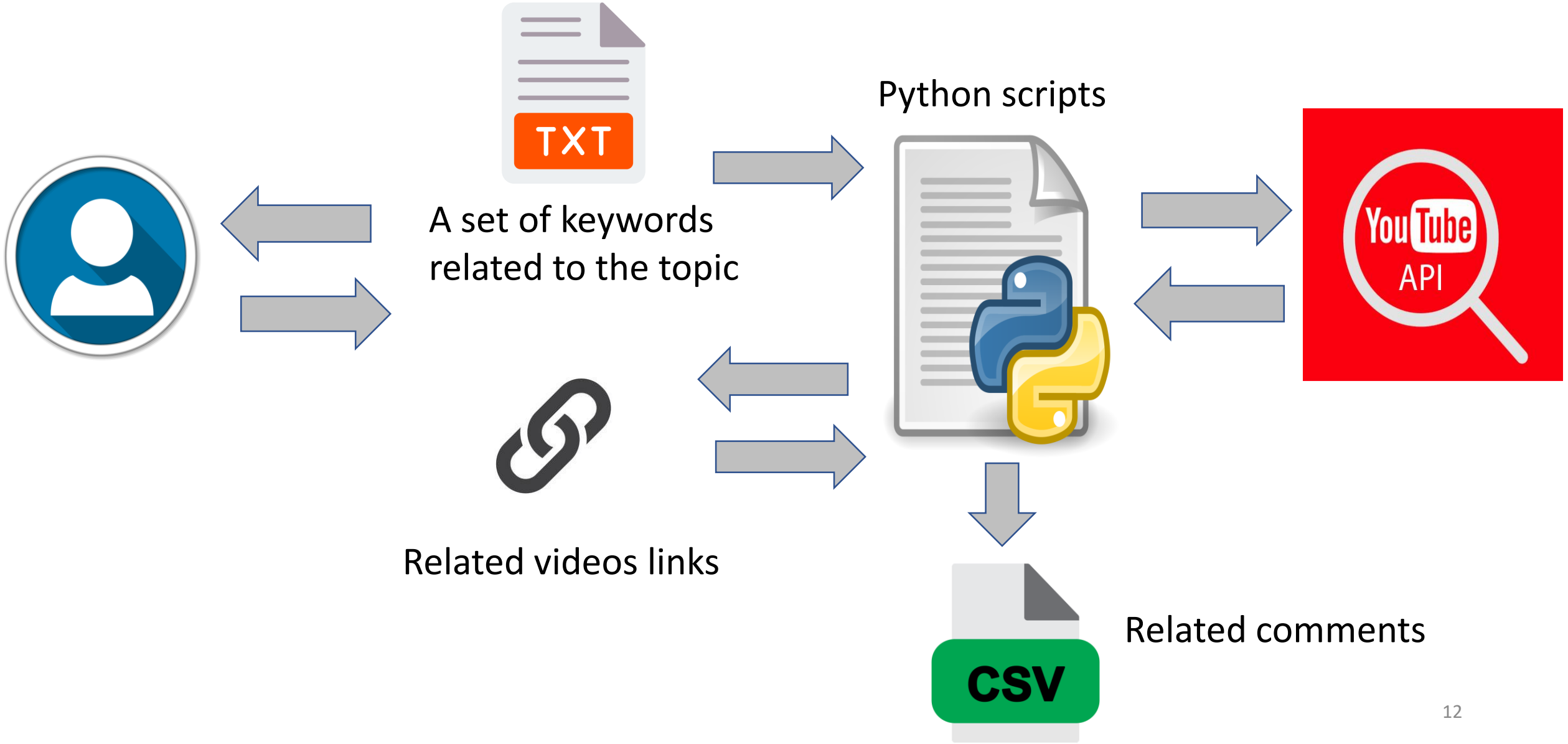
Twitter API

Python scripts



Relevant tweets and comments

MARASTA Corpus - Data Collection



MARASTA Corpus - Corpus Balancing

- ✓ For each region, we identified two controversial topics and collected sentences covering these topics.
- ✓ For each topic, we ensured that half of the sentences were in their respective region's dialect and the other half were in Modern Standard Arabic (MSA).
- ✓ Our corpus is well-balanced concerning both stance and dialect.

MARASTA Corpus - Annotation Guidelines

Definition of Stance

- The expression of being in favor of, against, or neutral towards a specific target.

Stance Labels:

- ”Pro” : expressing support or agreement
- ”Against”: expressing opposition or disagreement
- ”Neutral”: expressing neither support nor opposition

Annotation Process:

- Number of annotators per sentence, conflict resolution procedures, and quality control measures.

Examples and Edge Cases:

- Numerous examples illustrating each stance label, edge cases, and ambiguous instances to help annotators understand and consistently apply the guidelines.

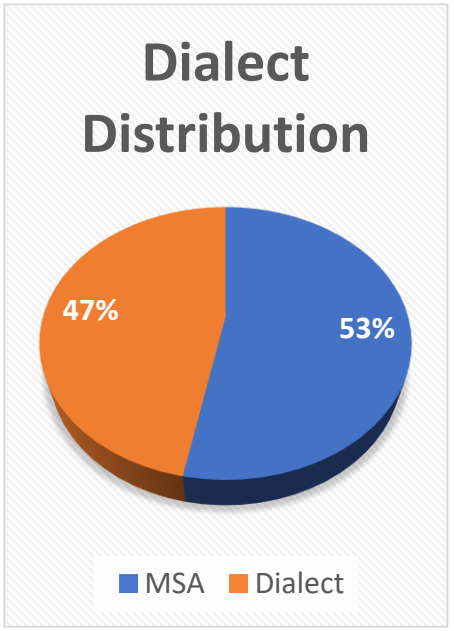
MARASTA Corpus - Corpus Annotation

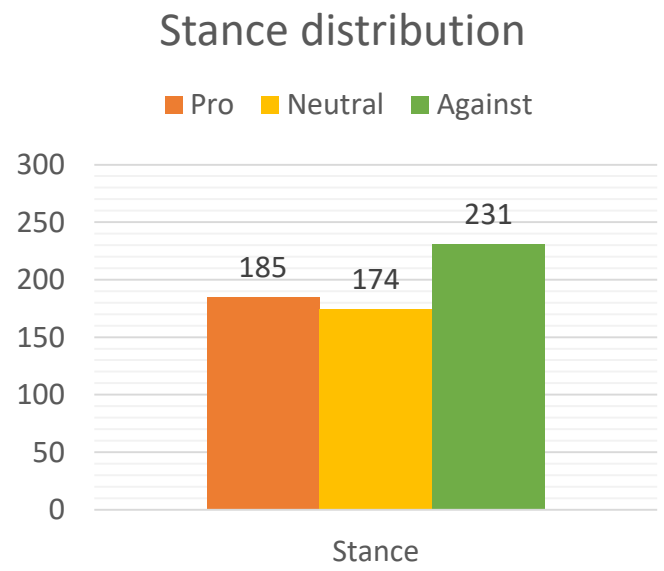
- The corpus was annotated by a team of skilled native Arabic speakers from the four regions we selected.
- Each sentence was annotated blindly by two annotators.
- After these two rounds of annotation, a script was run to detect **conflicts.**

- Two matching annotations would be considered as the final annotation. Otherwise, a discussion meeting should be done to agree on one annotation.
- If the three annotators agreed that a particular sentence was unclear or unrelated to the topic, it would be eliminated from the corpus.

MARASTA Corpus - Corpus statistics

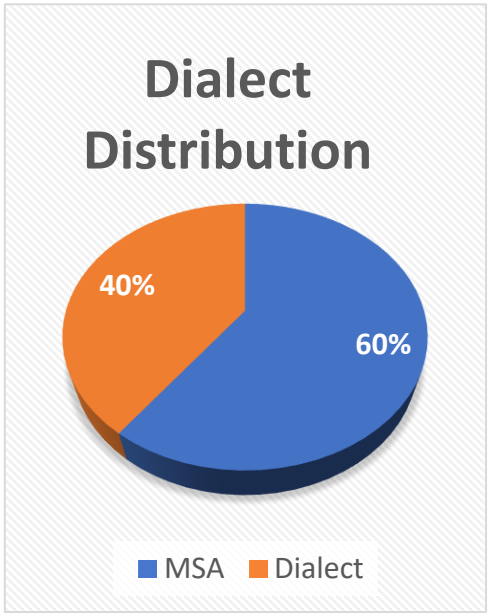
Region's Dialect: Maghreb

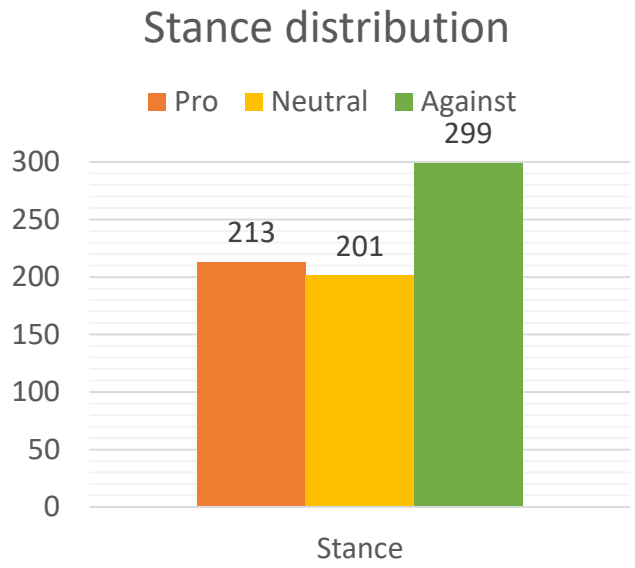




Total: 590

Illegal Immigration to Europe



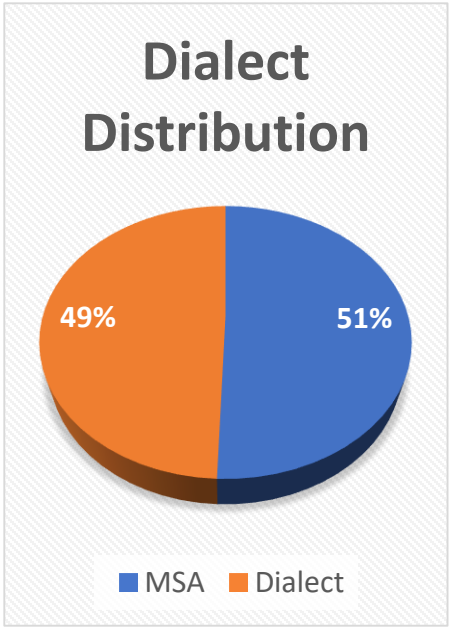


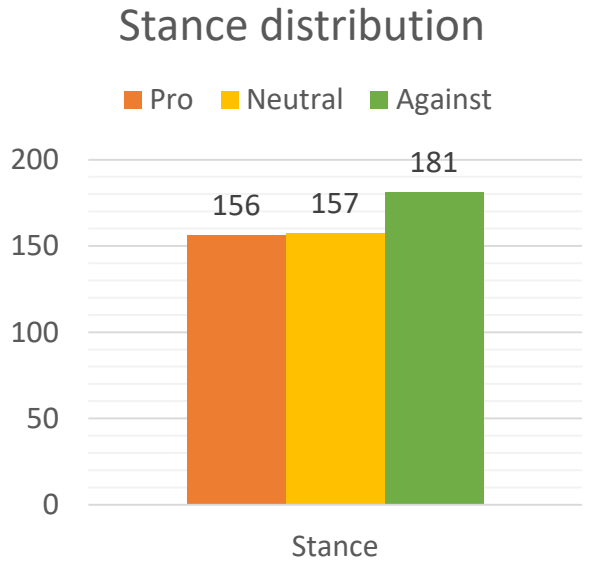
Total: 713

Tunisian General Labor Union

MARASTA Corpus - Corpus statistics

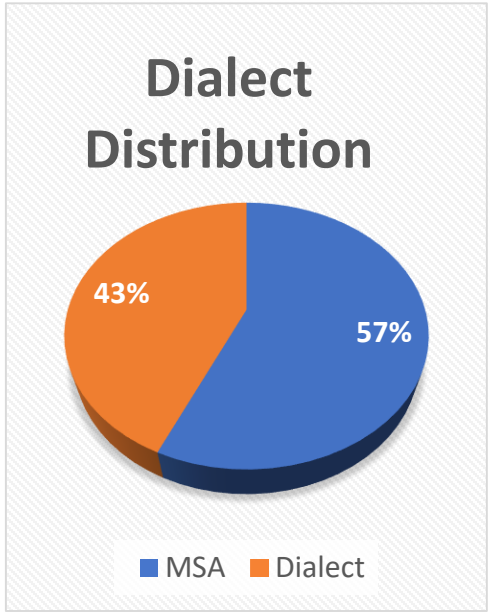
Region's Dialect: Egyptian

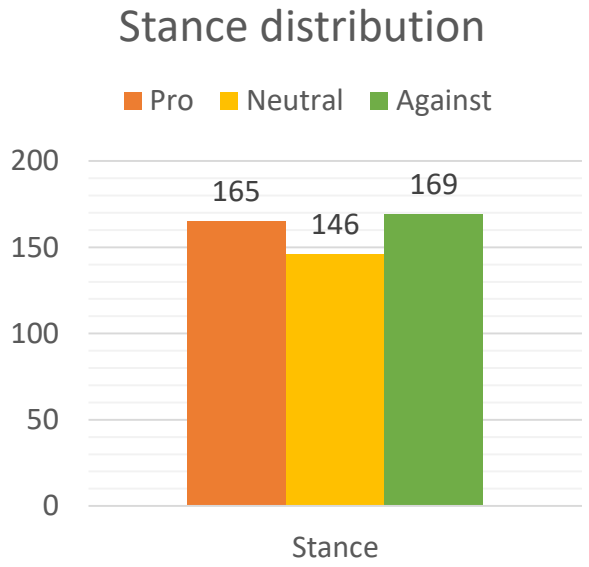




Total: 494

New Administrative Capital in Egypt



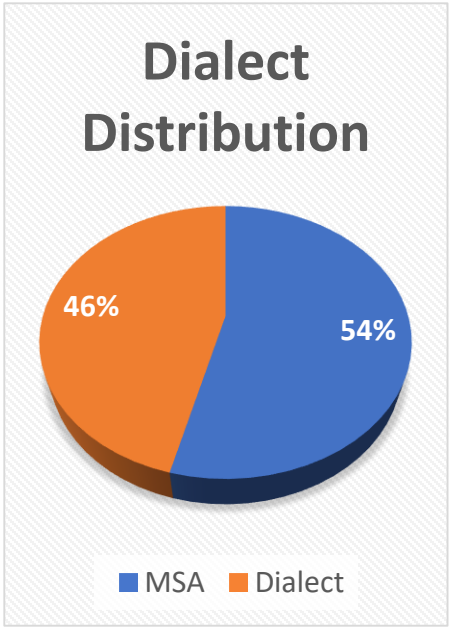


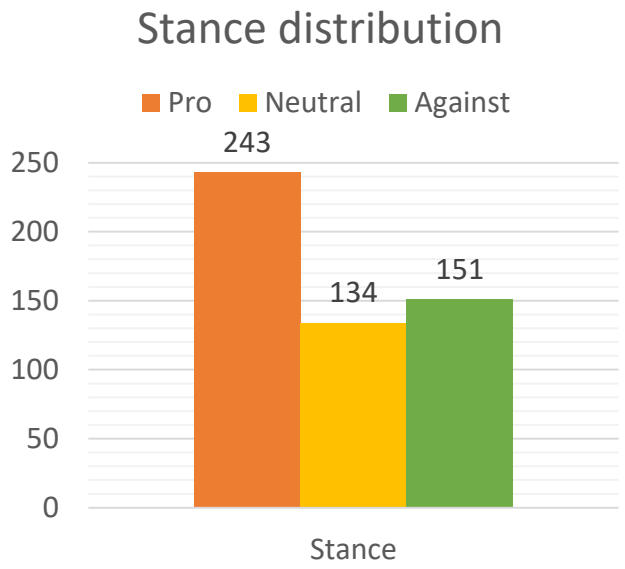
Total: 480

Egypt's 2013 Political Transition

MARASTA Corpus - Corpus statistics

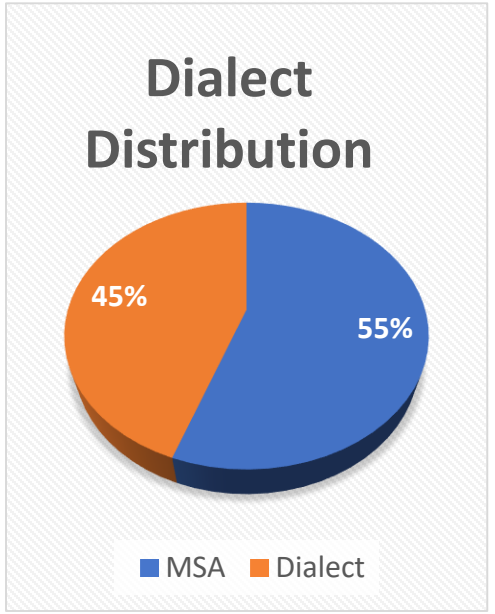
Region's Dialect: Gulf

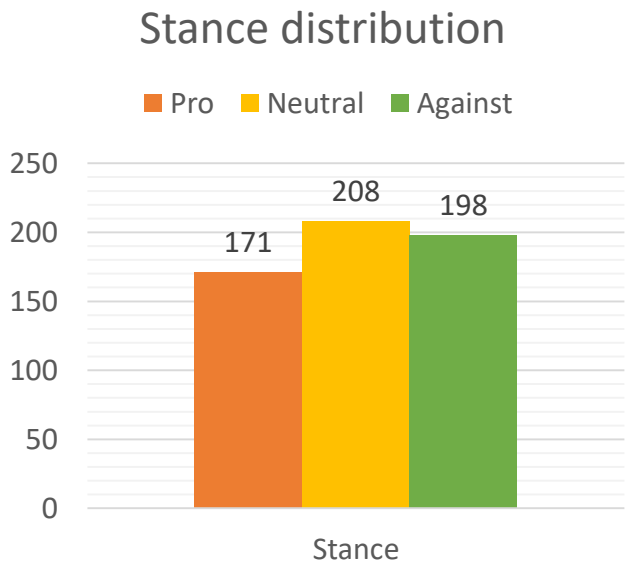




Total: 528

Fifa World Cup 2022



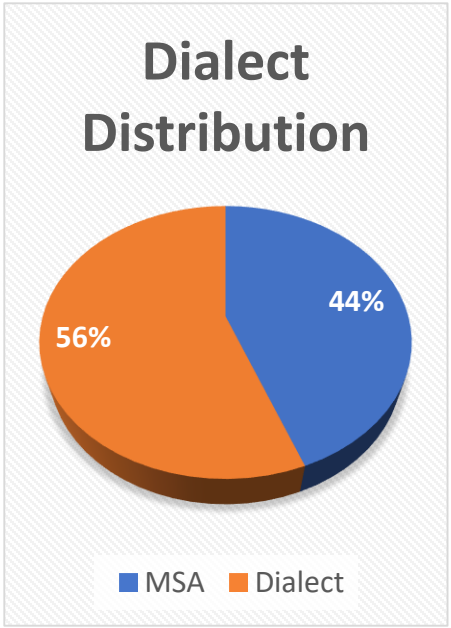


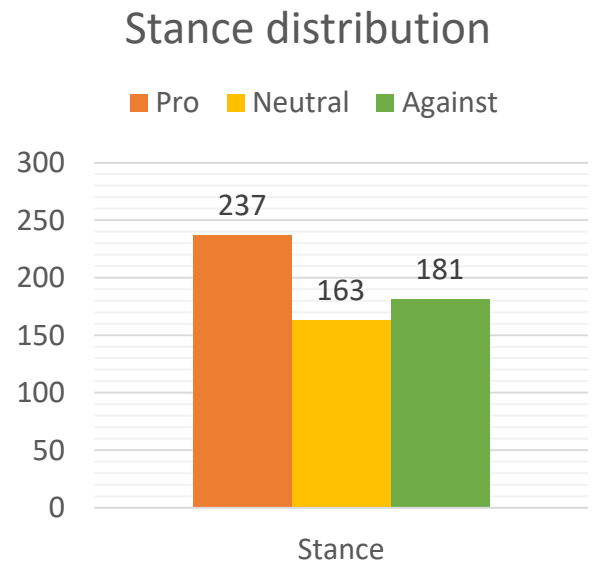
Total: 577

Normalization With Israel

MARASTA Corpus - Corpus statistics

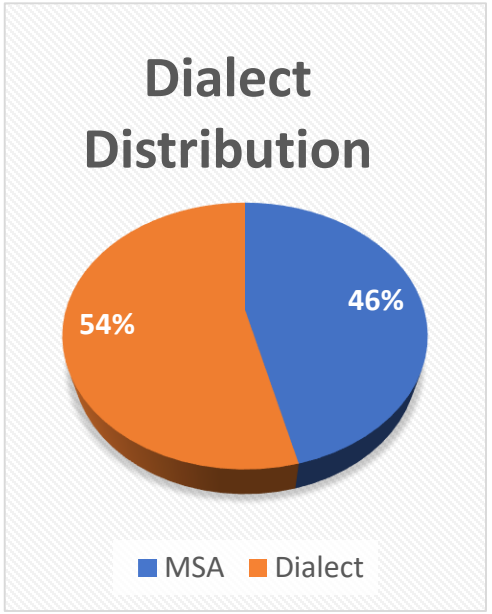
Region's Dialect: Levant

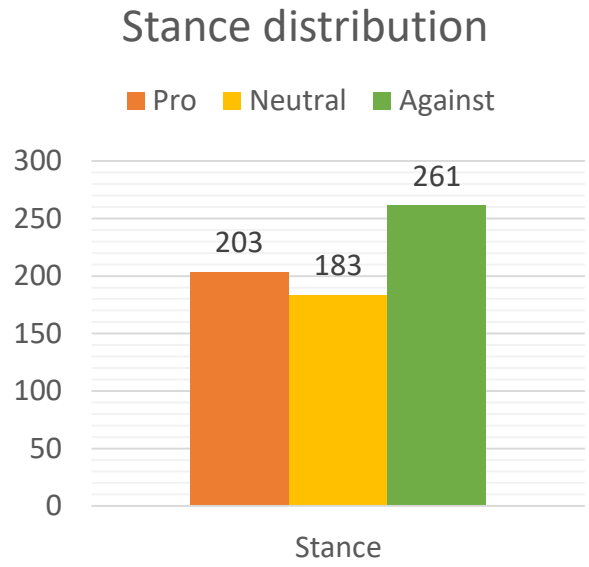




Total: 628

Presence of Refugees in Jordan





Total: 647

Feminism and Women's Rights

Experiments - Model Setup

- Training set: 3,492 samples , Testing set: 1,165 samples.
- Model Input: Topic name and related sentence
- Model Output: Classification label indicating the sentence's stance toward the topic

Pre-processing:

removing URLs, emails, stop words, punctuation, and non-Arabic characters

Features Extraction:

unigram-vectorizer, bigram-vectorizer, trigram-vectorizer, and tfidfvectorizer.

ML classification:

Support Vector Machine , Logistic Regression , Random Forest, and Decision Tree, and the Multi-Layer Perceptron

Experiments - Model Results

	Unigram		Bigram		Trigram		TF-IDF	
Classifier	***F1-score***	***Accu-racy***	***F1-score***	***Accu-racy***	***F1-score***	***Accu-racy***	***F1-score***	***Accu-racy***
Support Vector Machine	0.61	0.62	0.41	0.44	0.37	0.41	0.63	0.64
Logistic Regression	0.64	0.64	0.56	0.56	0.44	0.45	0.66	0.66
Random Forest	0.64	0.64	0.51	0.53	0.41	0.42	0.62	0.62
Decision Tree	0.56	0.56	0.53	0.54	0.42	0.43	0.52	0.52
Multi-Layer Perceptron	0.66	0.66	0.58	0.59	0.41	0.42	0.66	0.66

Experiments - Model Results

	Unigram		Bigram		Trigram		TF-IDF	
Classifier	***F1-score***	***Accu-racy***	***F1-score***	***Accu-racy***	***F1-score***	***Accu-racy***	***F1-score***	***Accu-racy***
Support Vector Machine	0.61	0.62	0.41	0.44	0.37	0.41	0.63	0.64
Logistic Regression	0.64	0.64	0.56	0.56	0.44	0.45	0.66	0.66
Random Forest	0.64	0.64	0.51	0.53	0.41	0.42	0.62	0.62
Decision Tree	0.56	0.56	0.53	0.54	0.42	0.43	0.52	0.52
Multi-Layer Perceptron	**0.66**	**0.66**	0.58	0.59	0.41	0.42	**0.66**	**0.66**

Experiments

- We tested the performance of MAWQIF dataset using the MLP model with TF-IDF features extraction.
- MARASTA and MAWQIF datasets were separately trained and tested using the same model.
- unlike MAWQIF, our dataset demonstrates more balanced performance across all stance classes.

Class	Metric	Dataset	
		MARASTA	MAWQIF
Pro	Precision	0.69	0.78
	Recall	0.63	0.86
	F1-score	0.66	0.82
Against	Precision	0.67	0.64
	Recall	0.63	0.63
	F1-score	0.65	0.64
Neutral	Precision	**0.61**	**0.42**
	Recall	**0.73**	**0.19**
	F1-score	**0.66**	**0.26**
Overall Accuracy		0.66	0.73
Macro F1-score		0.66	0.71

Importance of addressing class label imbalances

Conclusion and Future work

- Introduction of a novel **cross-domain multidialectal Arabic stance corpus** covering four regions: Maghreb, Egypt, Levantine, and the Gulf.
- The corpus comprises over **4,500 sentences** categorized into **eight distinct topics** across all regions.
- Annotation process involved at least two rounds of annotation, with a third round added in case of conflicts.
- Machine learning experiments conducted on the stance corpus for stance detection task, using different classifiers with various feature combinations.
- **The Multi-Layer Perceptron** (MLP) classifier yielded the best results when using **Unigram** or **TF-IDF features**.

Conclusion and Future work

- **Balanced distribution** of sentences across classes contributed to **consistent model performance** across different stance labels, emphasizing the importance of distribution balance.

- **Future work** will focus on developing **tools** for stance detection in Arabic and their application in **real-world scenarios**.

Carnegie Mellon University Qatar

جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

الصندوق القطري لرعاية البحث العلمي
Qatar National Research Fund
Member of Qatar Foundation

Thank you !

LREC-COLING 2024