LREC-COLING 2024

# So Hateful! Building a Multi-Label Hate Speech Annotated Arabic Dataset

Wajdi Zaghouani , Hamdy Mubarak , Md. Rafiul Biswas
Hamad Bin Khalifa University, Qatar
Qatar Computing Research Center, HBKU, Qatar

مجلس قطر
للبحوث
والتطوير
والابتكار
QRDI COUNCIL

الصندوق القطري لرعاية البحث العلمي
Qatar National Research Fund
عضو في مؤسسة قطر
Member of Qatar Foundation

جامعة
حمد بن خليفة
HAMAD BIN KHALIFA
UNIVERSITY
عضو في مؤسسة قطر
Member of Qatar Foundation

# Outlines

- Background
- Our Contribution
- Data Annotation Guidelines
- Results
- Limitation
- Conclusion

# Background

**Social media revolutionized** peoples **thoughts, opinions,** and experiences

Information is shared through social media spread easily

However**, the downside is that negative information and opinions are also easily spread**

# Background

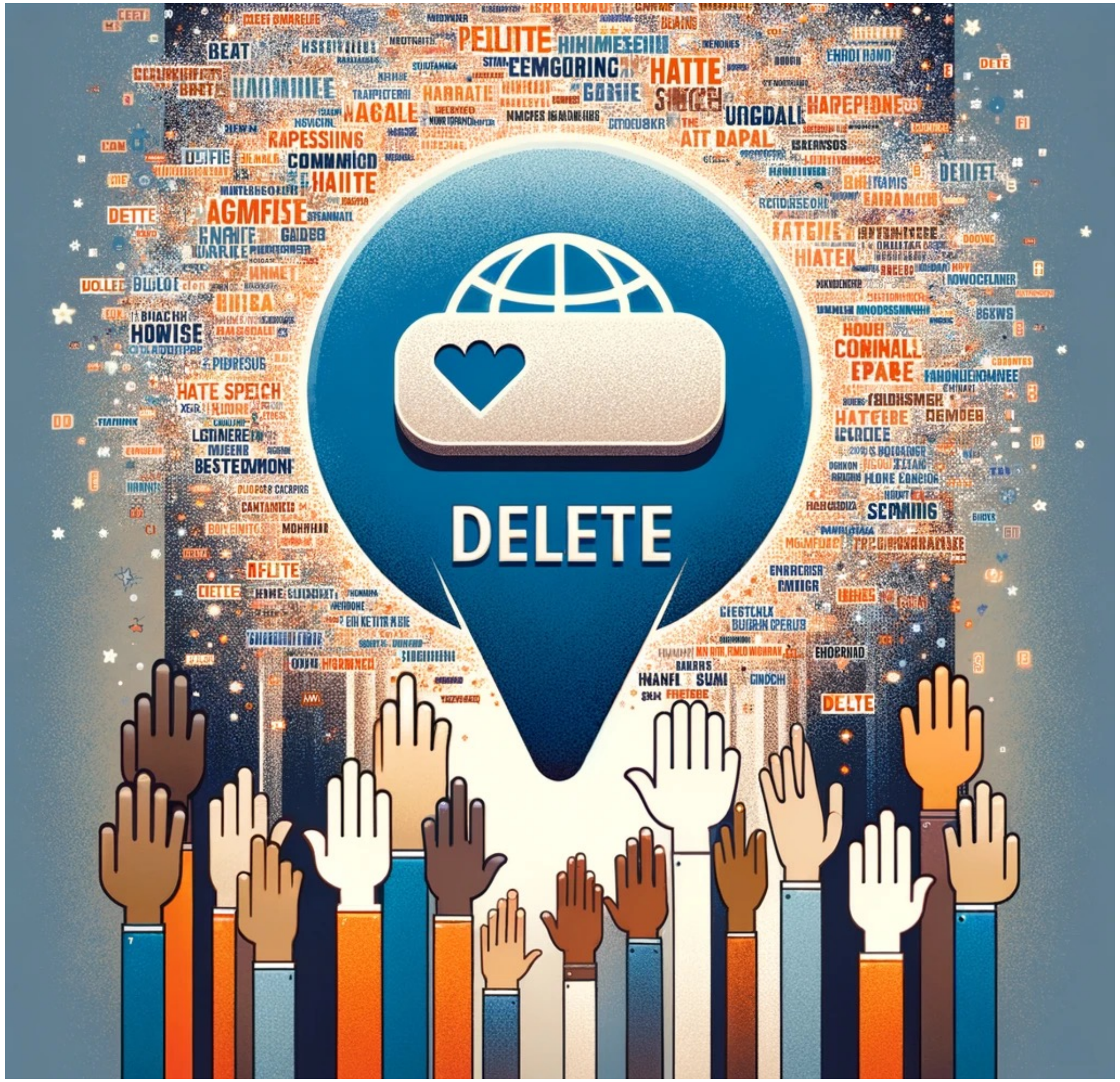
Hate speech is one negative form of expression that is prevalent on social media

Hate speech often arises with the emergence of events around the world

Hateful content is known to spread faster  than other content on social media

# Background

- No more Hate Speech
- The German government secured an agreement from social media platforms,  to delete all hate speech targeting refugees within 24 hours of its occurrence on the platform



# Motivation for this Research

Identifying and removing hate speech from social media is challenging  in Arabic
Language due to the diverse nature

The language has various dialects that differ from each other and from Modern Standard Arabic

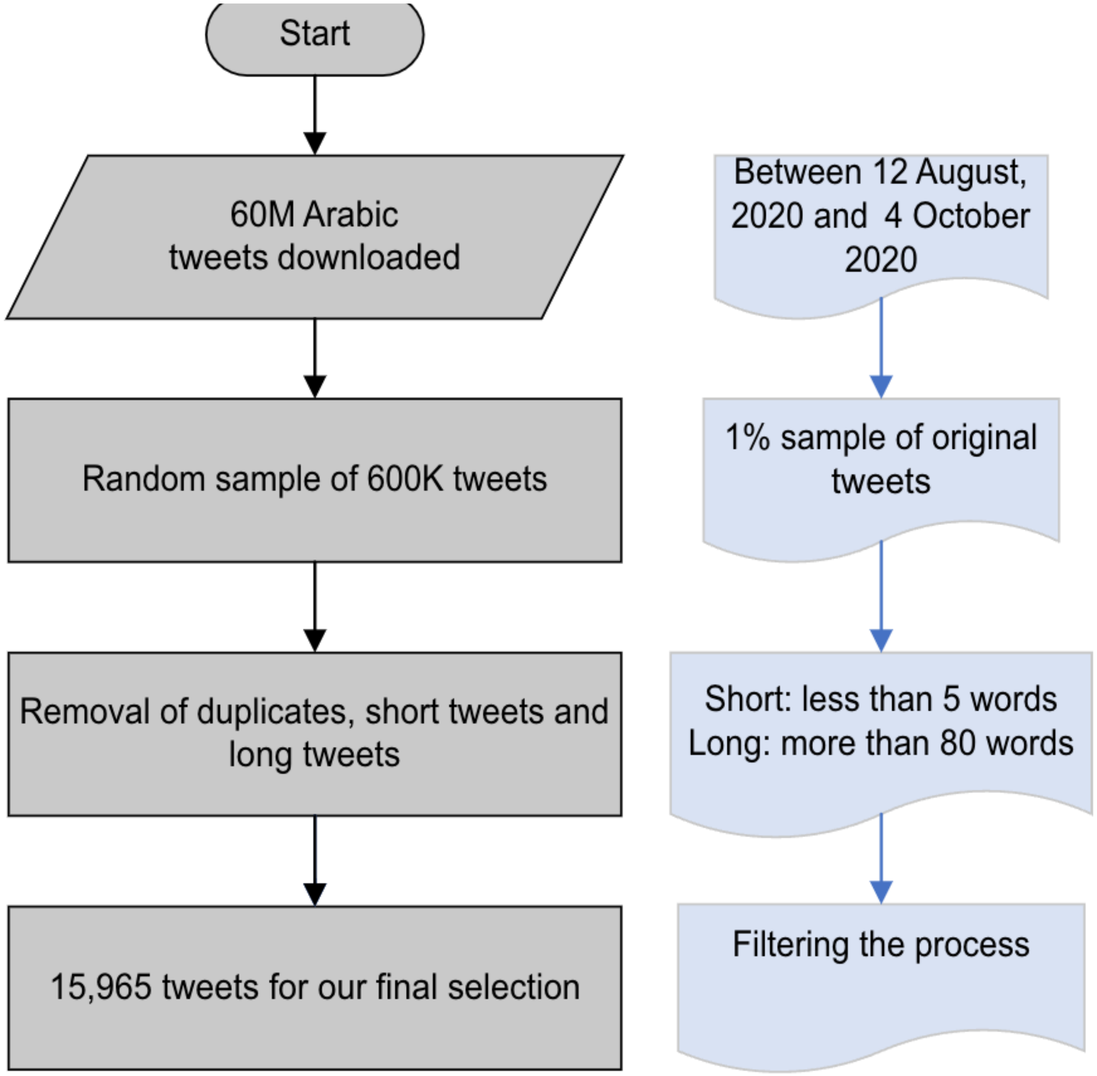
Various studies explored different approaches to annotating and detecting hate speech on Arabic Twitter

These studies collectively underscore the importance of developing  hate speech detection within the challenges of informal dialectal social media posts

# Our Contribution

- Created the largest multi-label, fine-grained Arabic hate speech dataset to date
- Our dataset is unique and versatile, with each tweet annotated with nine labels, such as sentiments, emotions, and valence, etc.
- Documented the dataset’s collection and guidelines and reproduceable for future projects
- Comprehensive corpus analysis of the dataset on the distinct features of Arabic hate speech discourse
- Carried out experiments with several classification techniques

# Data Collection


Start
60M Arabic tweets downloaded
Random sample of 600K tweets
Removal of duplicates, short tweets and long tweets
15,965 tweets for our final selection
Between 12 August, 2020 and 4 October 2020
1% sample of original tweets
Short: less than 5 words
Long: more than 80 words
Filtering the process

# Data annotation

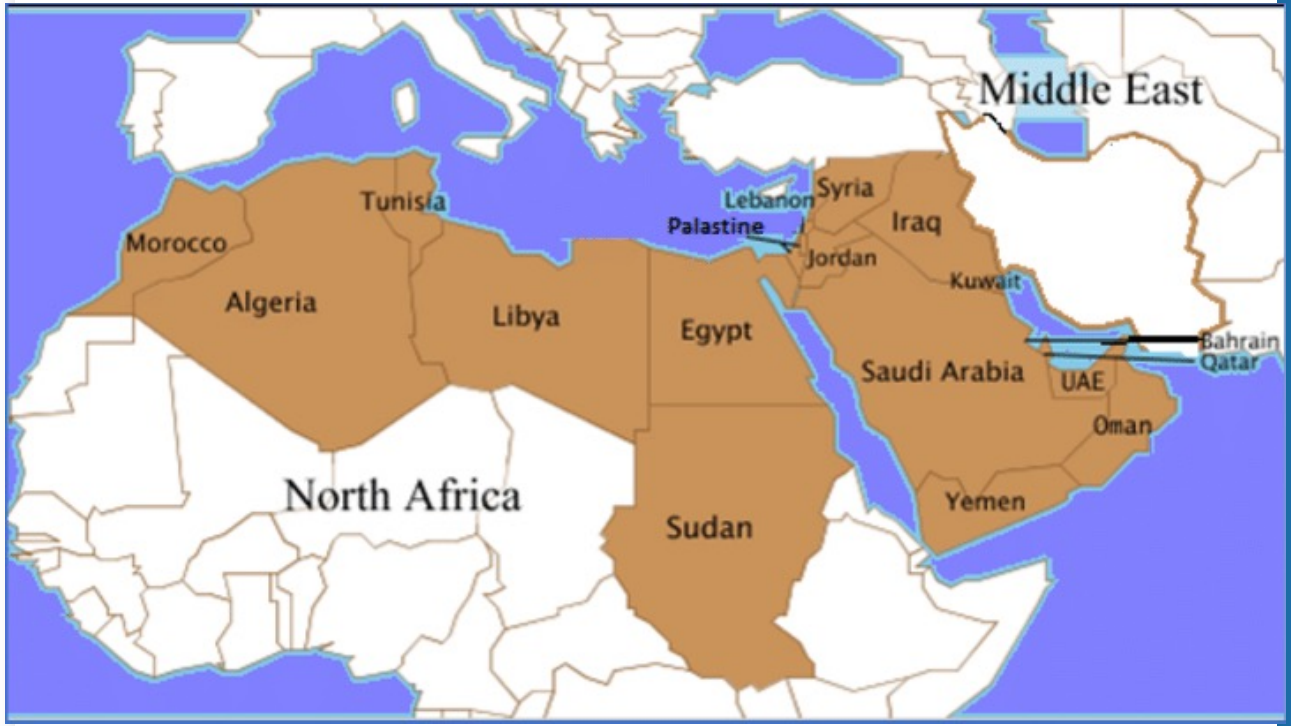
- We chose multi-label dataset and focused not only on offensive discourse and hate  speech
- Rather, we asked the annotators to label the tweets for the 13 categories

# Data annotation categories

| Categories | Subcategories |
|---|---|
| **Q1. Emotions** | choosing from 12 options like anger, anticipation, sadness etc. or neutral |
| **Q2. Emotion intensity** | no, small, moderate or large amount |
| **Q3. Sentiment** | very positive to very negative or neutral/mixed |
| **Q4. Offensive content** | whether tweet contains offensive language and if directed at a target |
| **Q4.1 Hate speech type** | individual, group, other entity |
| **Q4.2 Hate speech target** | race, religion, ideology, gender, social class |
| **Q4.3 Vulgarity** | whether offensive tweet contains profanity |
| **Q4.4 Violence** | if offensive tweet promotes violence |
| **Q5. Effect** | whether tweet is positive/inspiring or negative |
| **Q6.Sarcasm/irony** | whether directed at a target |
| **Q7 Humor** | not funny, somewhat funny or very funny |
| **Q8. Factuality** | if tweet contains verifiable information and is important |
| **Q9. Spam** | annoying advertising or requests |

# Annotators and training

- Location: Middle East and North Africa
- Language: Arabic
- Tweet Evaluation: Initially 16 Annotators
- Tweet Annotation: 1 to 5 annotators per tweet based on their ability to understand the dialect
- Training: Every two-three weeks trained to understand of guidelines, procedures, complex concepts of Arabic tweet



# Revision and Production

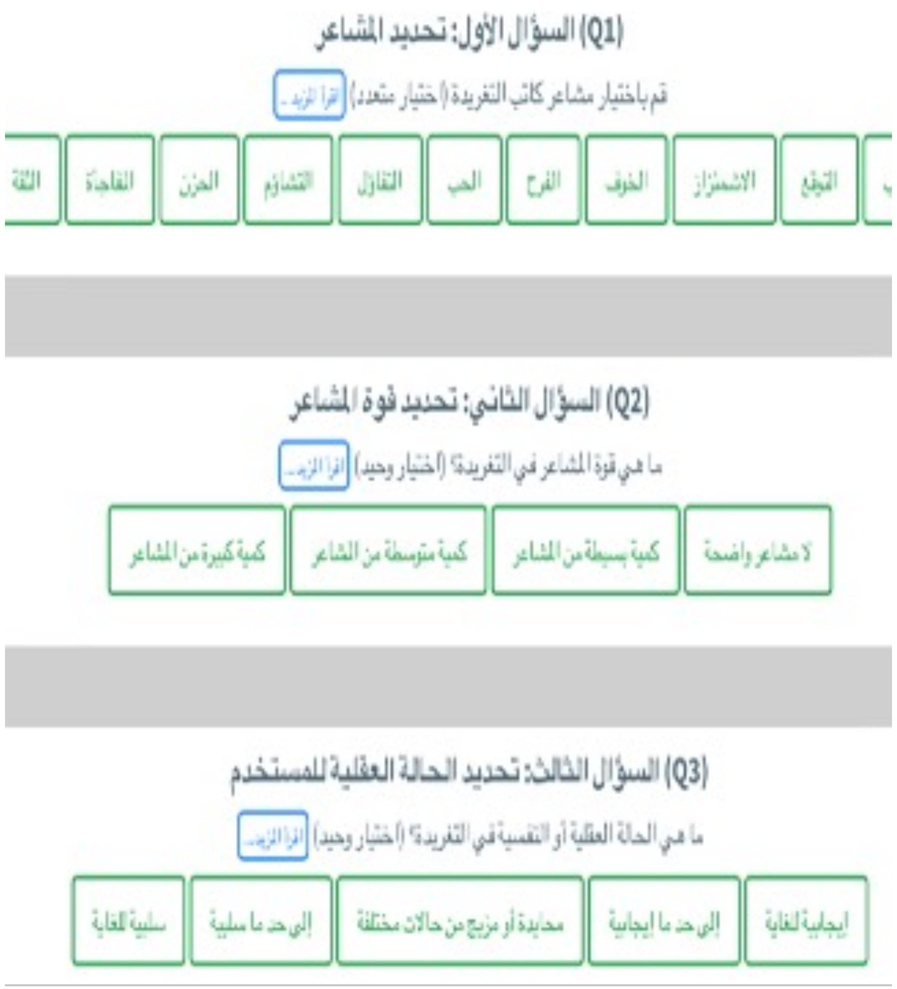
- Revision: Manager
    - analyzed errors,
    - unresolved cases,
    - feedback
    - updated guidelines

  to maximize quality, consistency, and consensus in annotation decisions

Production: Annotators met regularly but usually worked independently

# Annotation Interface

- MicroMappers is an online annotation management tool
- A screenshot from the Arabic version of the annotation interface (Showing the first three questions



(Q1) السؤال الأول: تحديد المشاعر

قم باختيار مشاعر كاتب التغريدة (اختيار متعدد)

(Q2) السؤال الثاني: تحديد قوة المشاعر

ما هي قوة المشاعر في التغريدة؟ (اختيار وحيد)

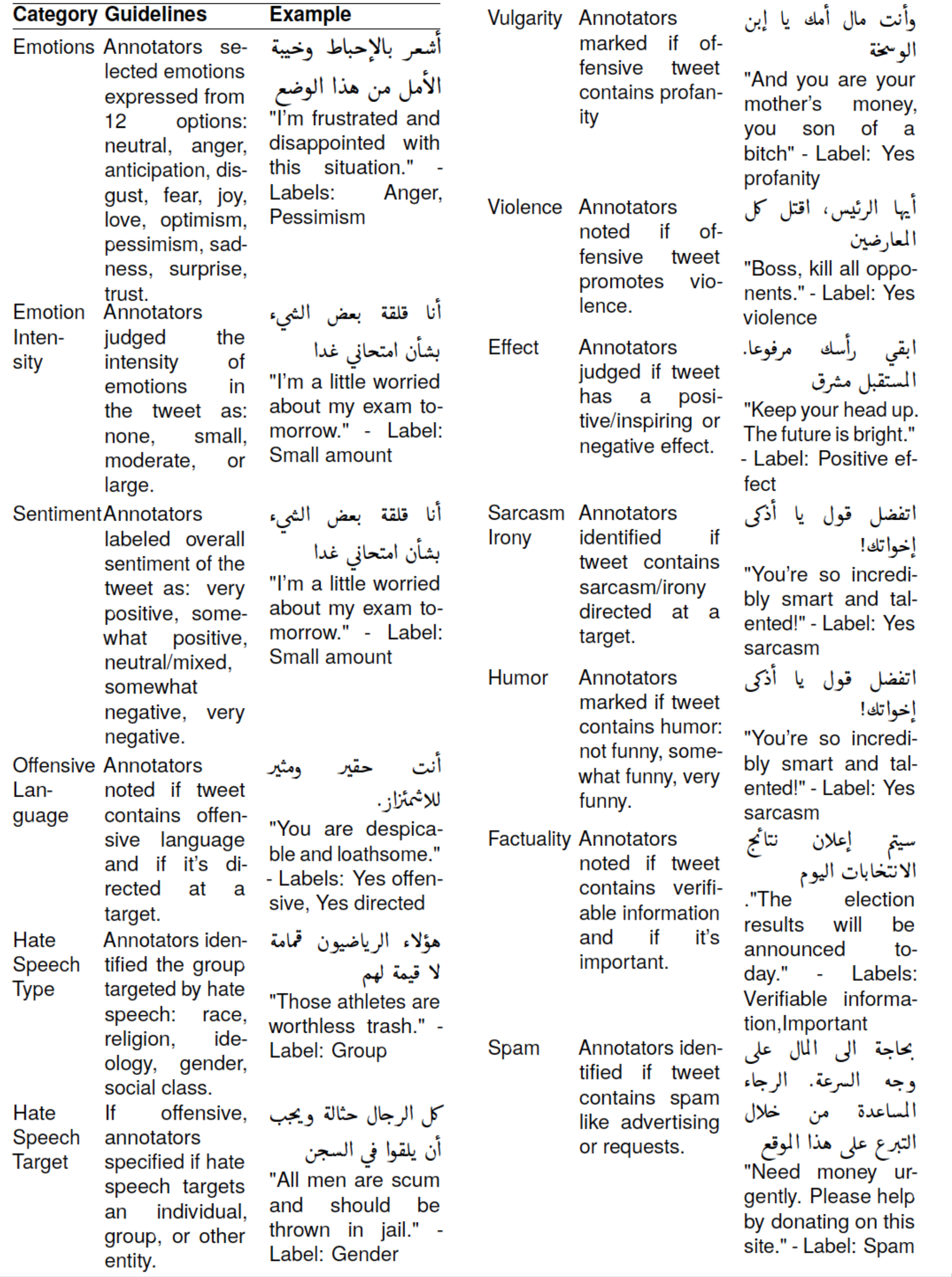
(Q3) السؤال الثالث: تحديد الحالة العقلية للمستخدم

ما هي الحالة العقلية أو النفسية في التغريدة؟ (اختيار وحيد)

# Annotation Guidelines

| Category | Guidelines | Example |
|---|---|---|
| **Emotions** | Annotators selected emotions expressed from 12 options: neutral, anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust. | أشعر بالإحباط وخيبة الأمل من هذا الوضع<br>"I'm frustrated and disappointed with this situation." - Labels: Anger, Pessimism |
| **Emotion Intensity** | Annotators judged the intensity of emotions in the tweet as: none, small, moderate or large. | أنا قلقة بعض الشيء بشأن امتحاني غدا<br>"I'm a little worried about my exam tomorrow." - Label: Small amount |

# Annotation Guidelines



| Category | Guidelines | Example |
|---|---|---|
| Emotions | Annotators selected emotions expressed from 12 options: neutral, anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust. | أشعر بالإحباط وخيبة الأمل من هذا الوضع "I'm frustrated and disappointed with this situation." - Labels: Anger, Pessimism |
| Emotion Intensity | Annotators judged the intensity of emotions in the tweet as: none, small, moderate, or large. | أنا قلقة بعض الشيء بشأن امتحاني غدا "I'm a little worried about my exam tomorrow." - Label: Small amount |
| Sentiment | Annotators labeled overall sentiment of the tweet as: very positive, somewhat positive, neutral/mixed, somewhat negative, very negative. | أنا قلقة بعض الشيء بشأن امتحاني غدا "I'm a little worried about my exam tomorrow." - Label: Small amount |
| Offensive Language | Annotators noted if tweet contains offensive language and if it's directed at a target. | أنت حقير ومثير للاشمئزاز. "You are despicable and loathsome." - Labels: Yes offensive, Yes directed |
| Hate Speech Type | Annotators identified the group targeted by hate speech: race, religion, ideology, gender, social class. | هؤلاء الرياضيون قمامة لا قيمة لهم "Those athletes are worthless trash." - Label: Group |
| Hate Speech Target | If offensive, annotators specified if hate speech targets an individual, group, or other entity. | كل الرجال حثالة ويجب أن يلقوا في السجن "All men are scum and should be thrown in jail." - Label: Gender |
| Vulgarity | Annotators marked if offensive tweet contains profanity | وأنت مال أمك يا إبن الوسخة "And you are your mother's money, you son of a bitch" - Label: Yes profanity |
| Violence | Annotators noted if offensive tweet promotes violence. | أيها الرئيس، اقتل كل المعارضين "Boss, kill all opponents." - Label: Yes violence |
| Effect | Annotators judged if tweet has a positive/inspiring or negative effect. | ابقي رأسك مرفوعا. المستقبل مشرق "Keep your head up. The future is bright." - Label: Positive effect |
| Sarcasm Irony | Annotators identified if tweet contains sarcasm/irony directed at a target. | اتفضل قول يا أذكى إخواتك! "You're so incredibly smart and talented!" - Label: Yes sarcasm |
| Humor | Annotators marked if tweet contains humor: not funny, somewhat funny, very funny. | اتفضل قول يا أذكى إخواتك! "You're so incredibly smart and talented!" - Label: Yes sarcasm |
| Factuality | Annotators noted if tweet contains verifiable information and if it's important. | سيتم إعلان نتائج الانتخابات اليوم ."The election results will be announced today." - Labels: Verifiable information,Important |
| Spam | Annotators identified if tweet contains spam like advertising or requests. | بحاجة الى المال على وجه السرعة. الرجاء المساعدة من خلال التبرع على هذا الموقع "Need money urgently. Please help by donating on this site." - Label: Spam |

# Annotation Analysis

| | N | % |
|---|---|---|
| **Offensive language?** | | |
| Yes | 2793 | 17.5% |
| No | 13171 | 82.5% |
| **Directed?** | | |
| Yes, directed | 2348 | 84.7% |
| No, not directed | 445 | 15.93% |
| **Target2** | | |
| Individual | 963 | 41.01% |
| People with common features | 1090 | 46.42% |
| Organization, company, situation, or topic | 299 | 13.9% |
| **What do they have in common?** | | |
| Ideology, politics, sports | 747 | 68.5% |
| Class, social status, or profession | 83 | 7.6% |
| Religion or sect | 144 | 13.3% |
| Gender | 43 | 3.9% |
| Origin, race, or nationality | 382 | 3.5% |
| **Obscene language?** | | |
| Yes | 874 | 31.29% |
| No | 1919 | 69.71% |
| **Advocates for violence?** | | |
| Yes | 201 | 7.1% |
| No | 2592 | 92.8% |

# Sentiment Results from Manual Annotation

| Effect of the Tweet on the Reader’s Wellbeing/Annotator’s Sentiment | N | % |
|---:|---:|---:|
| Frustrating | 4237 | 26.54 |
| Motivating | 2637 | 15.51 |
| Neither frustrating nor motivating | 9089 | 56.93 |

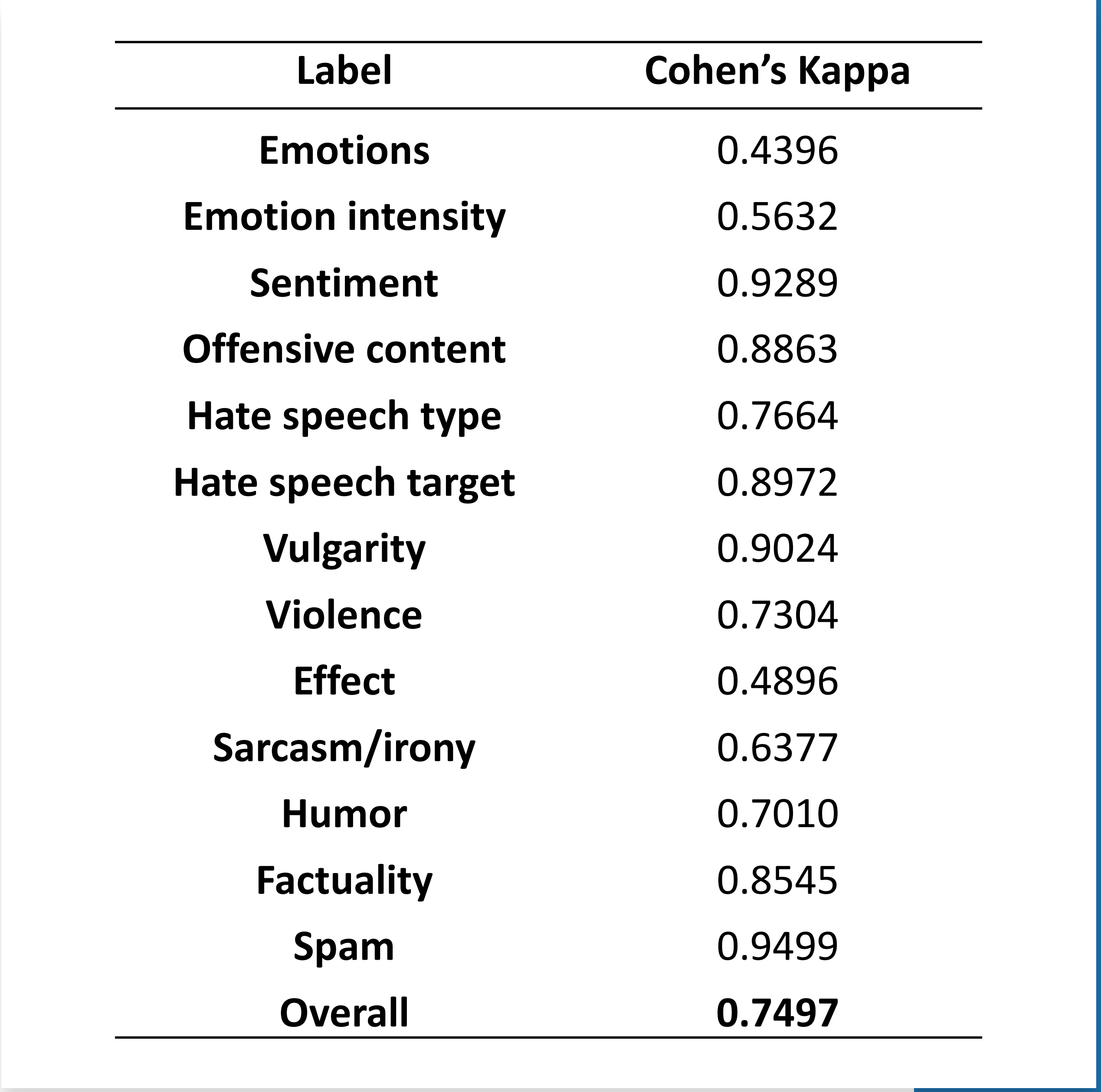
# Fact-Checking

| Fact-Checking | N | % |
|---|---|---|
| No information | 10190 | 63.83% |
| Contains information, but not verifiable | 2987 | 18.71% |
| Contains information that is verifiable | 2787 | 17.45% |
| **Important to the public?** | | |
| Yes, important | 1594 | 57.19% |
| No, not important | 981 | 35.2% |

# Data Annotation Categories

| Data Annotation Categories | N |
|---|---|
| **Emotion (Pessimism,sadness,confidence, joy and others)** | 12301 |
| **Emotion Intensity (Small, large, average amount of feelings and others)** | 9075 |
| **Sarcasm, irony, and ridicule (yes, no)** | 9036 |
| **Humor, joking (Yes, but not funny and others)** | 9257 |
| **Spam Detection (Yes, specific product or service, and others)** | 9004 |
| **Valence (positive, negative. Neutral, and others)** | 9059 |

# Annotation Evaluation



| Label | Cohen’s Kappa |
|---|---|
| **Emotions** | 0.4396 |
| **Emotion intensity** | 0.5632 |
| **Sentiment** | 0.9289 |
| **Offensive content** | 0.8863 |
| **Hate speech type** | 0.7664 |
| **Hate speech target** | 0.8972 |
| **Vulgarity** | 0.9024 |
| **Violence** | 0.7304 |
| **Effect** | 0.4896 |
| **Sarcasm/irony** | 0.6377 |
| **Humor** | 0.7010 |
| **Factuality** | 0.8545 |
| **Spam** | 0.9499 |
| **Overall** | **0.7497** |

# Example of Annotation Disagreement

| Example tweet | **Annotator Disagreement** |
|---|---|
| الله يبارك فيك حبيبي ي احمد عقبال عندك يارب ❤️❤️😍<br>God bless you [in response to “congratulations”], my dear Ahmed. I wish you the same ❤️❤️😍 | One annotator opted for a "neutral" emotional classification, and the two others selected "joy#optimism#confidence" and "optimism#confidence". |

# Offensive Language Detection Using Machine and Deep Learning

| Model | Label | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|---|
| **LR** | Yes | 0.58 | 0.34 | 0.43 | 0.84 |
| | No | 0.87 | 0.95 | 0.91 | |
| **RF** | Yes | 0.70 | 0.25 | 0.37 | 0.85 |
| | No | 0.86 | 0.98 | 0.91 | |
| **GB** | Yes | 0.81 | 0.14 | 0.23 | 0.84 |
| | No | 0.84 | 0.99 | 0.91 | |
| **SVM** | Yes | 0.49 | 0.40 | 0.44 | 0.82 |
| | No | 0.87 | 0.91 | 0.89 | |
| **DT** | Yes | 0.42 | 0.38 | 0.40 | 0.79 |
| | No | 0.87 | 0.88 | 0.88 | |
| **Ara-bert** | All | 0.49 | 1.00 | 0.65 | 0.82 |

# HateSpeech Detection Using Machine and Deep Learning

| Model | Label | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|---|
| **LR** | Yes | 0.38 | 0.14 | 0.20 | 0.93 |
| | No | 0.94 | 0.98 | 0.96 | |
| **RF** | Yes | 0.47 | 0.05 | 0.09 | 0.93 |
| | No | 0.93 | 1.00 | 0.96 | |
| **GB** | Yes | 0.41 | 0.04 | 0.07 | 0.93 |
| | No | 0.93 | 1.00 | 0.96 | |
| **SVM** | Yes | 0.27 | 0.24 | 0.25 | 0.90 |
| | No | 0.94 | 0.95 | 0.95 | |
| **DT** | Yes | 0.22 | 0.19 | 0.21 | 0.90 |
| | No | 0.94 | 0.95 | 0.95 | |
| **Arabert** | All | 0.57 | 0.47 | 0.66 | 0.83 |

# Limitations

The annotators originated from certain Arabic-speaking regions. This could introduce labeling biases based on regional dialects and interpretations of hate speech.

Hate speech involves inherent subjectivity which may have impacted the annotation accuracy

As Twitter was the sole data source, the findings might not reflect other social media platforms

Potential sampling bias may occur as we randomly sampled on our dataset

Majority were non-hate class which causes class imbalance,  inflate model performance metrics

# Conclusion

Multiple annotators from various Arabic-speaking regions ensure diversity on dataset annotation

Improvement and enhancement of the dataset is ongoing as well as refinement of the guidelines and annotations

This project serves as a valuable resource for researchers and practitioners in the field of Arabic language processing and analysis

Thank you

Q & A