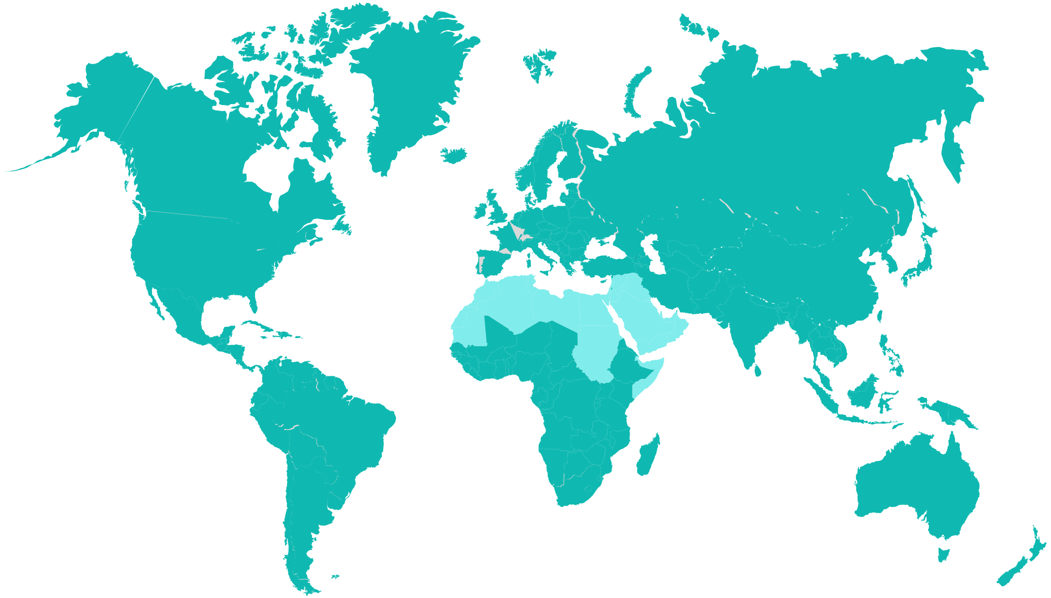
Arabic Sentiment Analysis

By: Khadejaa Saad Alshehri



The Arabic language is one of the Semitic languages used as official or co-official in around 20 countries. It is the sixth most spoken language worldwide, with an estimated 274 speakers in millions

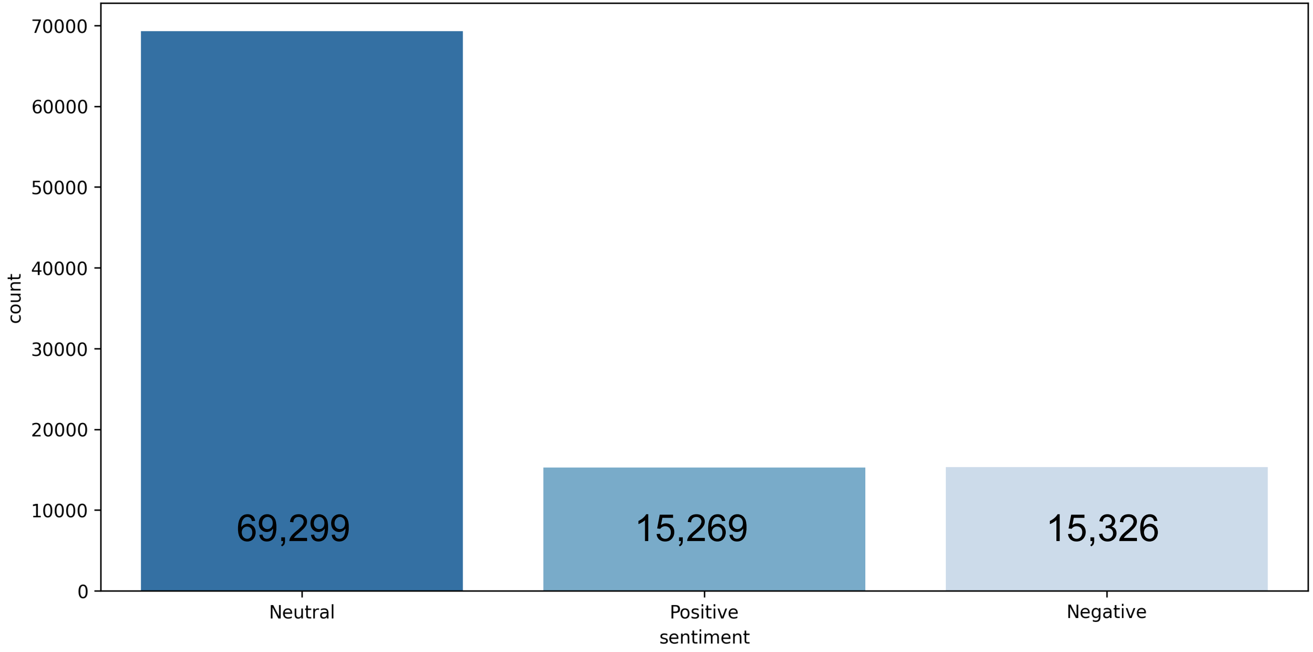
Twitter

Twitter provides a good channel for the public to share about products, personalities, events, our society etc. Even more interesting than these is the fact that people are telling about themselves. which make it a rich channel for sentiment analysis.

Data Description

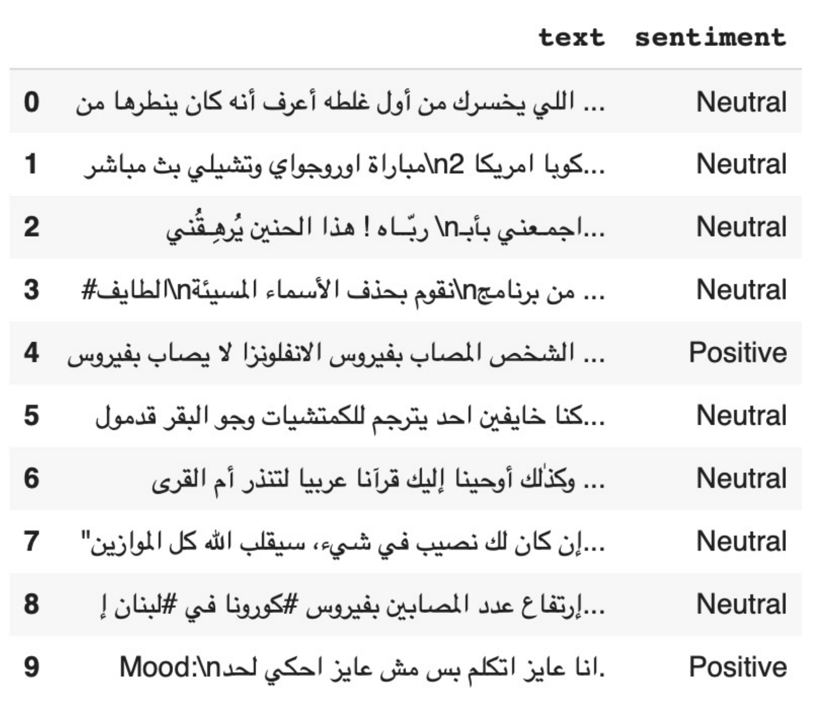
In this project, we used ASAD which consists of 95,000 annotated tweets, with three-class sentiments. ASAD has a total of 15,215 positive tweets, 15,267 negative tweets, and 64,518 neutral tweets.

Number of Samples



Cleaning the Dataset

- Encoded URL
- Remove punctuation.
- Remove html tags.
- Remove tashkeel, tatweel
- Normalize some letters
- Remove stop words



	text	sentiment
0	اللي يخسرك من أول غلطه أعرف أنه كان ينطرها من ...	Neutral
1	كوبا امريكا 2\nمباراة اوروجواي وتشيلي بث مباشر...	Neutral
2	اجمـعني بأبـ\n ربّـاه ! هذا الحنين يُرهِـقُني...	Neutral
3	من برنامج\nنقوم بحذف الأسماء المسيئة\nالطايف# ...	Neutral
4	الشخص المصاب بفيروس الانفلونزا لا يصاب بفيروس ...	Positive
5	كنا خايفين احد يترجم للكمتشيات وجو البقر قدمول...	Neutral
6	وكذٰلك أوحينا إليك قرآنا عربيا لتنذر أم القرى ...	Neutral
7	إن كان لك نصيب في شيء، سيقلب الله كل الموازين"...	Neutral
8	إرتفاع عدد المصابين بفيروس #كورونا في #لبنان إ...	Neutral
9	Mood:\nانا عايز اتكلم بس مش عايز احكي لحد.	Positive

Models

Logistic Regression CountVectorizer	**Logistic Regression** TfidfVectorizer	**Random Forest Classifier** TfidfVectorizer
Support Vector Machine TfidfVectorizer	**BiLSTM** word2vec	**BERT** BERT

Models

Logistic Regression CountVectorizer accuracy 0.64	**Logistic Regression** TfidfVectorizer accuracy 0.69	**Random Forest Classifier** TfidfVectorizer accuracy 0.66
Support Vector Machine TfidfVectorizer accuracy 0.69	**BiLSTM** Word2vec accuracy 0.69	**BERT** BERT accuracy 0.69

For Future Works

- Experiment with different pre-prossece.
  - Study effect of removing emoji/encoding it.
  - Study effect of removing #, @, URL.
- Experiment with hand-crafted features
  - Binary features if it content emoji or not if it vulgar words or not, content opinion word or not
  - Length of the tweet in terms of words and char
- Experiment with different BERT models.

Thanks!

Does anyone have any questions?