Arabic Sentiment Analysis:
Comparison of the ML Algorithms

By : Amirah Alshamrani

- **Motivation**
- Saudi Arabia is among the countries with the highest number of Twitter users among its online population.
- Development of models to track social behavior that fit the culture.
- The Machine Learning (ML) algorithms is a powerful technique for sentiment analysis.
- **Questions**
- Algorithm needs to evaluate by applying it to Arabic data set to choice best model and ML algorithms preferences.
- Our body of knowledge about the discourse of controversial social and cultural issues in Saudi Arabia is still limited.
- **Data**
- Data was collected from the from Twitter by using Crimson Hexagon.
- Each tweet was annotated as “positive” or “negative”.

Mythology

Prepare Data for Machine Learning Algorithm

- **Feature Extraction (or vectorization)**

Text need to be encoded or transformed as numerical values to be used as input to a machine learning algorithm, (Tf-idf) term-frequency times inverse document-frequency.

- **Removing stop words**

The package nltk.corpus stopwords has Arabic language as parameter

stopwords_list = stopwords.words(’Arabic')

- **Stemming**

Metrics Used to Evaluate Models

- **Accuracy**

Not sufficient to select a model for this problem.

- **Confusion Matrix**

Table that presents both the class distribution in the data and the classifiers predicted class distribution with a breakdown of error types

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Metrics Used to Evaluate Models

- **Precision**

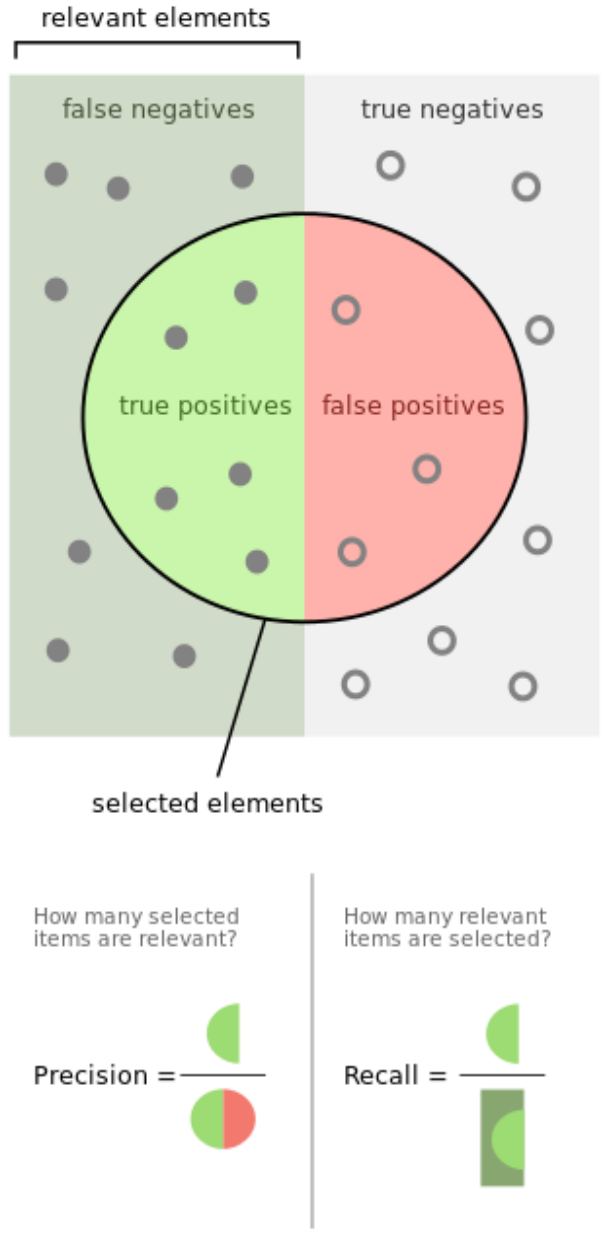
Number of true positive predictions divided by the total number of positive class values predicted.

- **Recall**

Number of true positive predictions divided by the number of positive class values in the test data.

- **F1 score**

Conveys the balance between the precision and the recall



Metrics Used to Evaluate Models

Neither precision (less false positive or proportion of stance classified as positive being actually positive) nor recall (less false negative or proposition of positive opinions identified as positive) has higher priority or importance than the other.

Therefore, to convey that balance between these two measures we chose f1 as a metric to evaluate different models

Tuning the hyper-parameters and estimating models

- Gridsearch
- 10-fold cross validation
- Trying different combinations of features

Evaluating Classifiers

- Naive Bayes is known to be an effective classifier for text data, I started with the text classification using it solely with one feature: tweet.
- Combining text with other features using the Dataframe mapper.
- SVM performed the best .

Evaluating Classifiers

Table 1: Stance Classification Performance

Classifier	Accuracy	F1-Score
Logistic Regression	0.59	0.59
SVM	0.60	0.59
Naïve Bayes	0.57	0.51

Table 2: Gender Classification Performance

Classifier	Accuracy	F1-Score
Logistic Regression	0.57	0.45
SVM	0.60	0.49
Naïve Bayes	0.53	0.40

For both gender and stance, the accuracy, as well as F1 score were highest for SVM model. Followed by the logistic regression and lastly the Naïve Bayes model.

Evaluating Classifiers – Limitations

Limitations

- The manual mark-up required annotators to understand Arab and Saudi culture. Therefore, our dataset is comparatively small.
- Using Twitter and Crimson Hexagon as a data source and collection tool involves multiple types of potential sampling biases

شكرًا