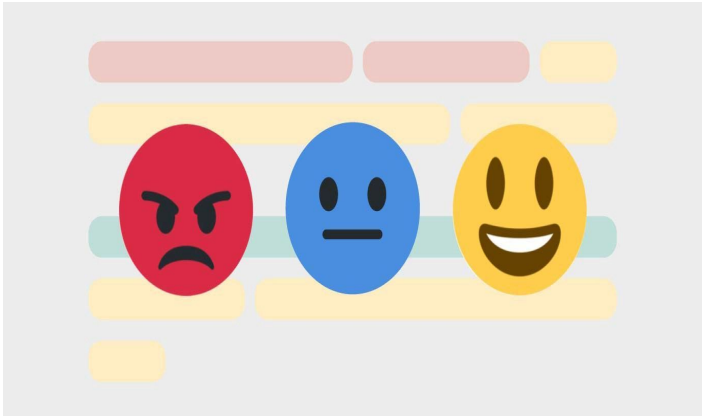
Sentiment Analysis For Low-resource African Languages using Twitter Dataset

by Ahmed Issaoui, Massa Coulibaly, and Ahmad Mohamad



Introduction & Motivation

- Task Description: What is sentiment classification?
- Why is it important?
- What are we doing new? Algerian Dialect

Data Description

- From where did we get the data?
- Some information about the data:
  - Low quantity (~1600)
  - Low quality (وعلاه جای مکرون)
  - Not balanced (54% , 25%, 21%)

Pre-processing

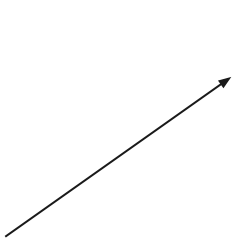
1. Removing @user and RT (Twitter tokens)
2. Tokenization
3. Removing stopwords
4. Removing punctuation
5. Removing numbers
6. Normalizing emojis
7. Removing empty string tokens

@user تبهليل ما بعد منتصف الليل😂😂😂 مرنكة أقسم بالله

مرنكة أقسم بالله 😂 تبهليل منتصف الليل

Data Augmentation

تبهليل هاذا



هاذا [MASK]

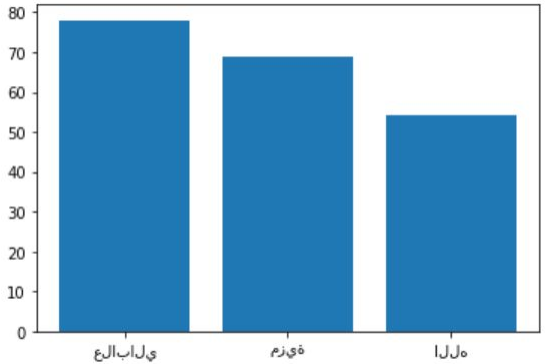
[MASK] تبهليل

شكون هاذا

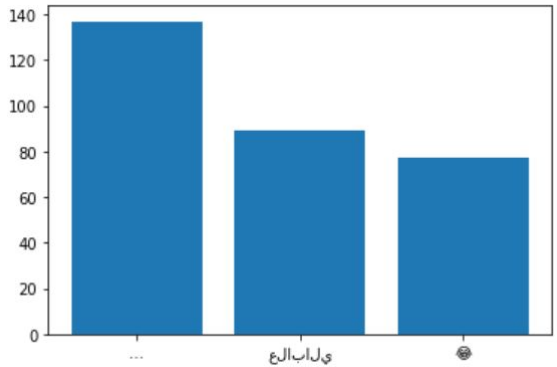
تبهليل شبعتونا

DziriBERT

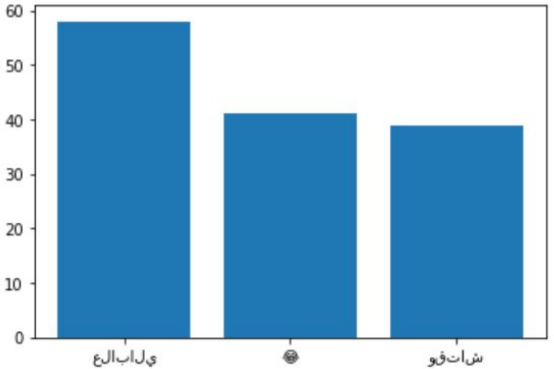
Data Analysis



Most Freq Positive Words



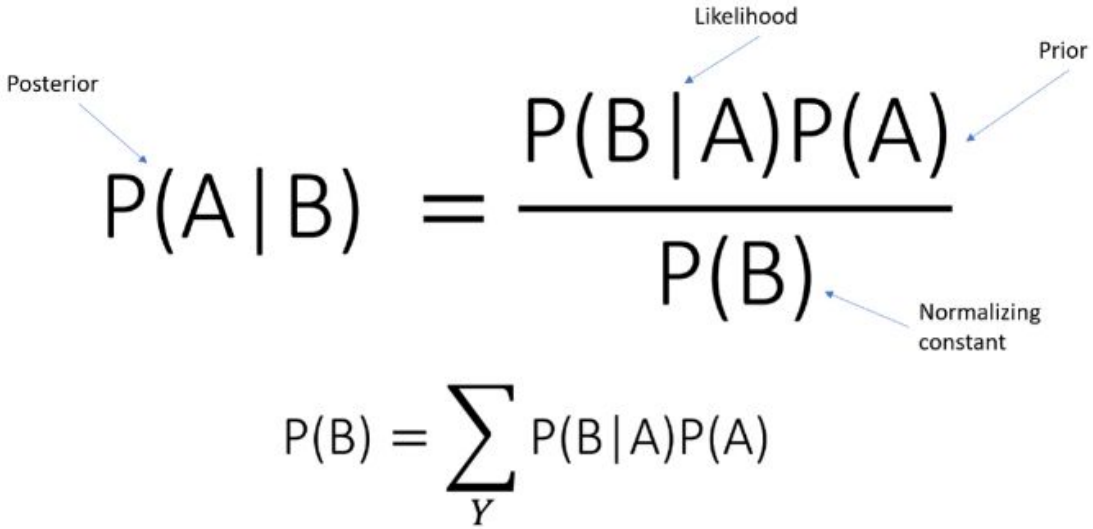
Most Freq Negative Words



Most Freq Neutral Words

Basic ML models

Naive Bayes

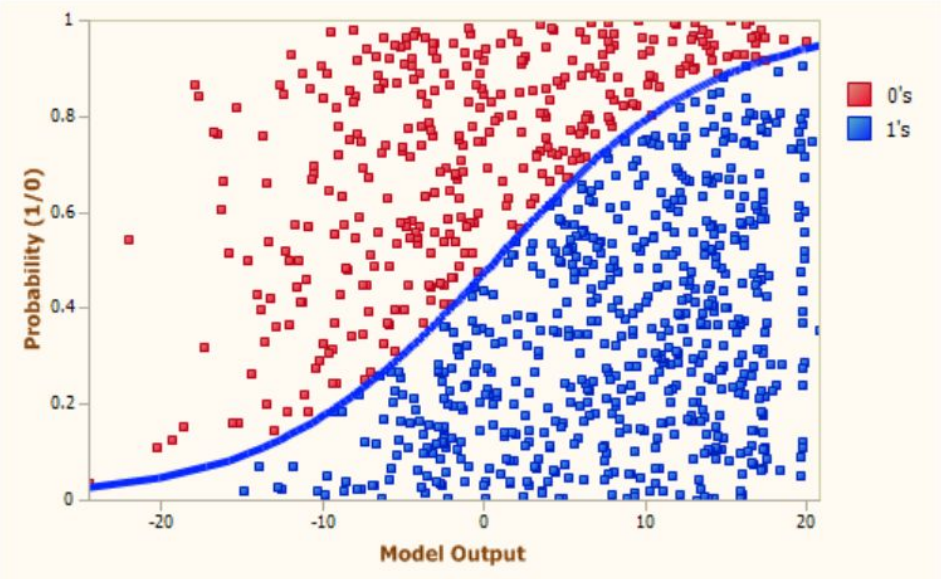


Results:

Non-Augmented: 51%

Augmented: 54%

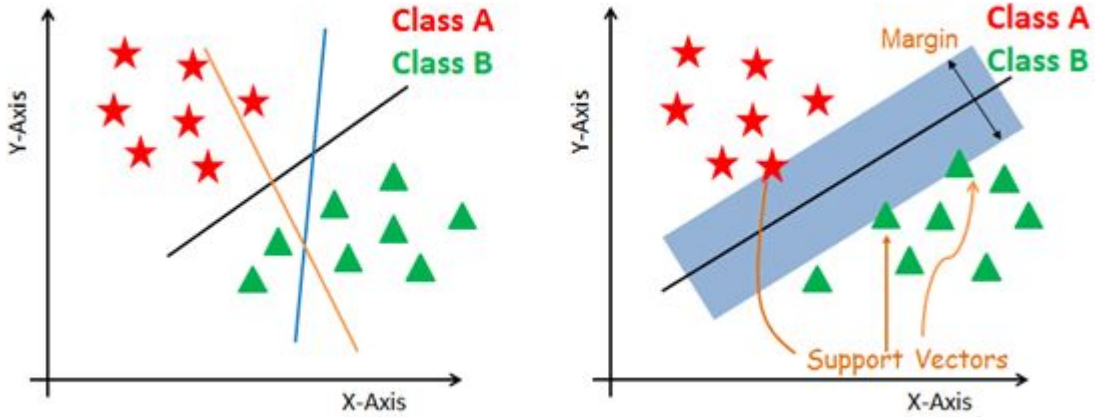
Logistic Regression



Results:

Non-Augmented: 48%

SVM

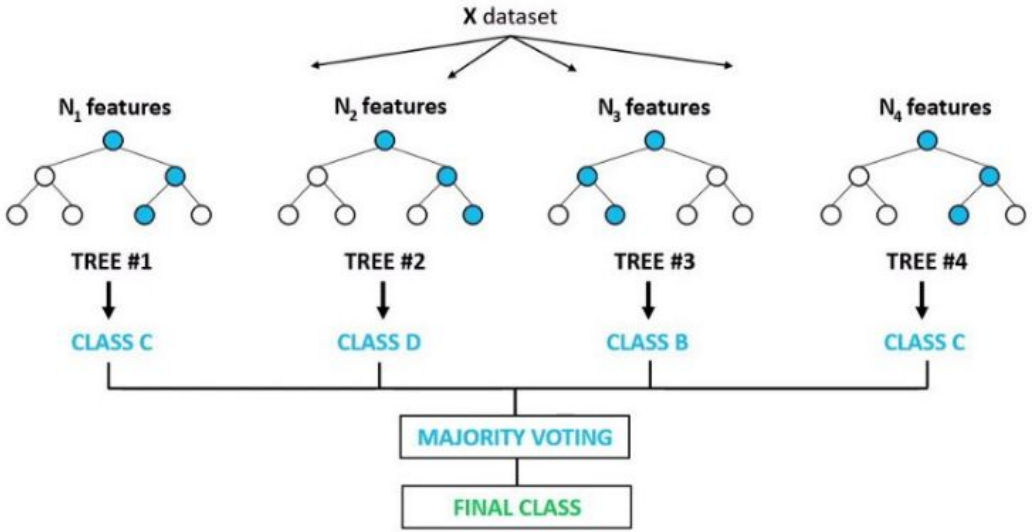


Results:

Non-Augmented: 37%

Augmented: 41%

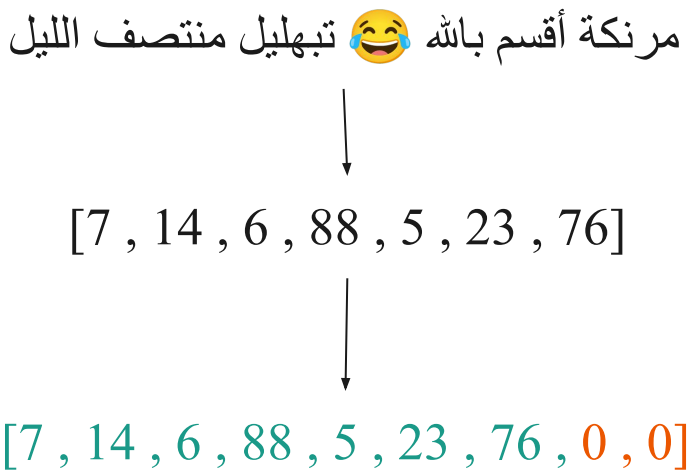
Random Forest



Results:

Non-Augmented: 46%

LSTM-based models: Vectorization & Padding



LSTM-based models: Layers

```
model=Sequential()
model.add(Embedding(total_words,embedding_size,embeddings_initializer=Constant(embedding_matrix),input_length=MAX_SEQUENCE_LENGTH,trainable=True
model.add(LSTM(68, dropout = 0.5))
model.add(Dense(3,activation='softmax'))
model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])
checkpoint = ModelCheckpoint(f"{project_dir}/best_model.hdf5", monitor='val_accuracy', verbose=1,save_best_only=True, mode='auto',save_weights_c
callback = tensorflow.keras.callbacks.EarlyStopping(monitor='loss', patience=4)
history= model.fit(x_train, to_categorical(y_train, num_classes=3), epochs=100,callbacks=[checkpoint,callback],validation_data=(x_test, to_categ
```

LSTM-based models: Results

- 61% accuracy
- +10% improvement on baseline models

LSTM-based + Aravec

bakrianoo/**aravec**

AraVec is a pre-trained distributed word representation (word embedding) open source project which aims to provide the Arabic NLP research...

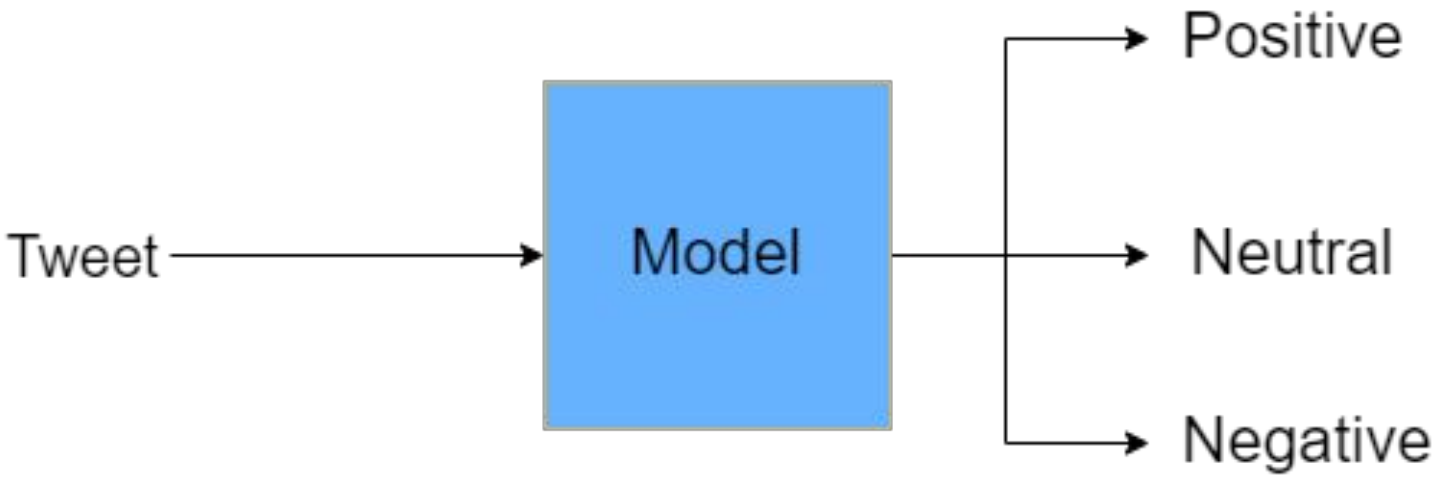
2 Contributors 3 Issues 314 Stars 73 Forks

```
model=Sequential()
model.add(Embedding(total_words,embedding_size,embeddings_initializer=Constant(embedding_matrix),input_length=MAX_SEQUENCE_LENGTH,trainable=True
model.add(LSTM(68, dropout = 0.5))
model.add(Dense(3,activation='softmax'))
model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])
checkpoint = ModelCheckpoint(f"{project_dir}/best_model.hdf5", monitor='val_accuracy', verbose=1,save_best_only=True, mode='auto',save_weights_
callback = tensorflow.keras.callbacks.EarlyStopping(monitor='loss', patience=4)
history= model.fit(x_train, to_categorical(y_train, num_classes=3), epochs=100,callbacks=[checkpoint,callback],validation_data=(x_test, to_categ
```

LSTM-based + Aravec: Results

- 63% accuracy
- +2% improvement on the basic LSTM model

New approach: 2-step Classification


Tweet
Model
Positive
Neutral
Negative

I`m going to uni 😭😭😡

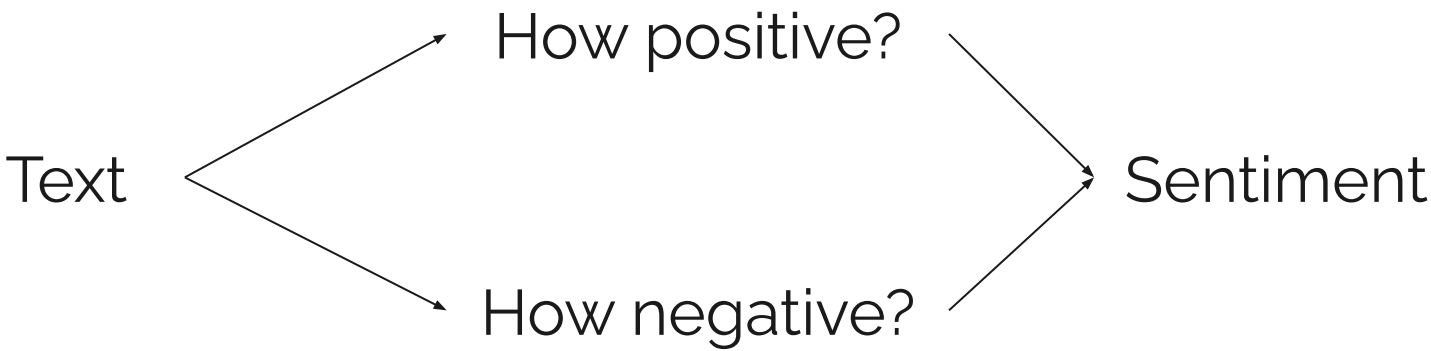
I`m going to uni 😊😊😊

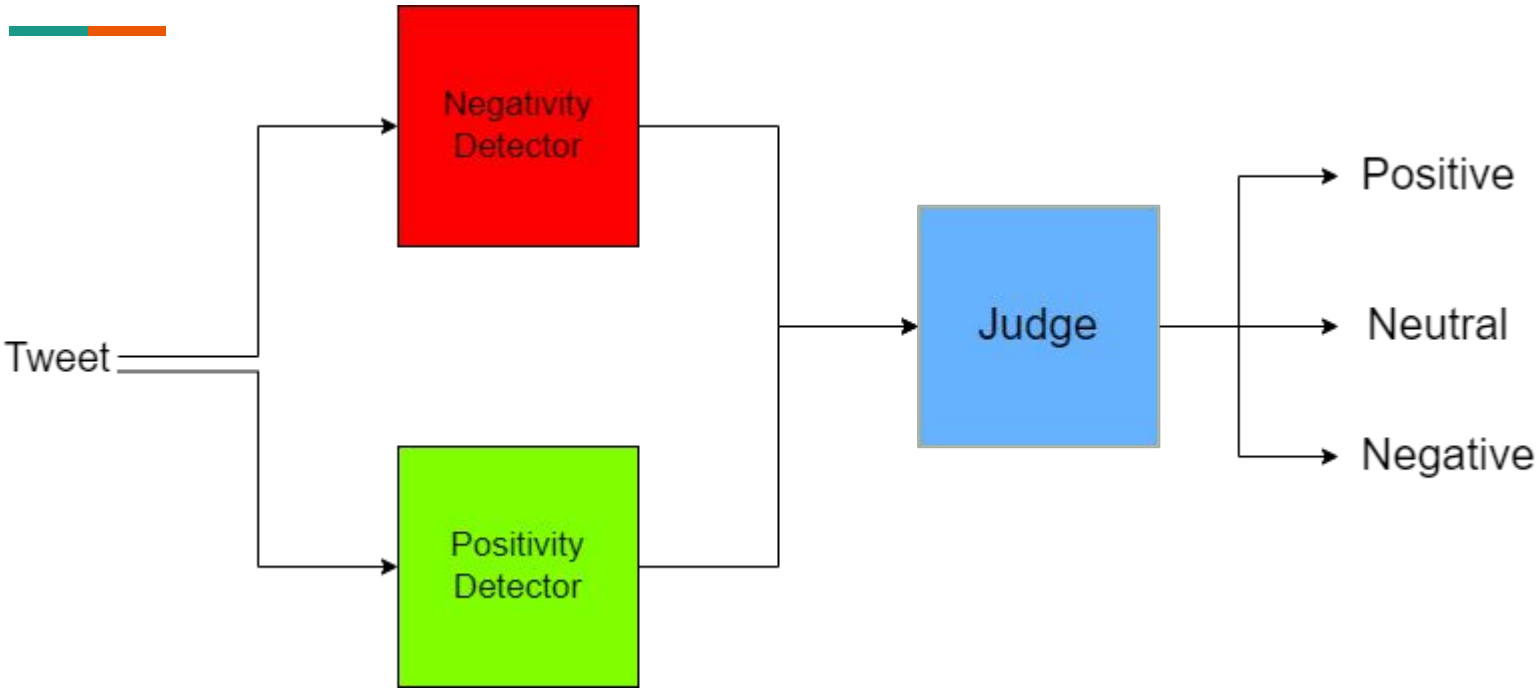
I`m going to uni

If there is no patterns that indicate neutrality in a sentence (neutrality itself is not detectable)

how are we (as humans) able to identify neutral sentences?

We do not directly identify neutrality?
We go through a 2-step classification process




Negativity Detector
Tweet
Positivity Detector
Judge
Positive
Neutral
Negative

Results

- 70% accuracy
- +20% improvement on baseline models
- 7% improvement on traditional approach (LSTM-based model)

Conclusion and Future Work

Better Stopwords

it is really hard to collect all stop words from the algerian dialect.

If we get access to these stop words then there will be more focus on the important information.

Better Data

To make a enhanced model, better quality and more data and will need to be collected.

The current data is very low quality and quantity but we tried to get the best accuracy possible.

Better Augmentation

Spend more time on making a more advanced data augmentation algorithm.

Our augmentation algorithm is very basic and needs more work if we are working with arabic dialects.