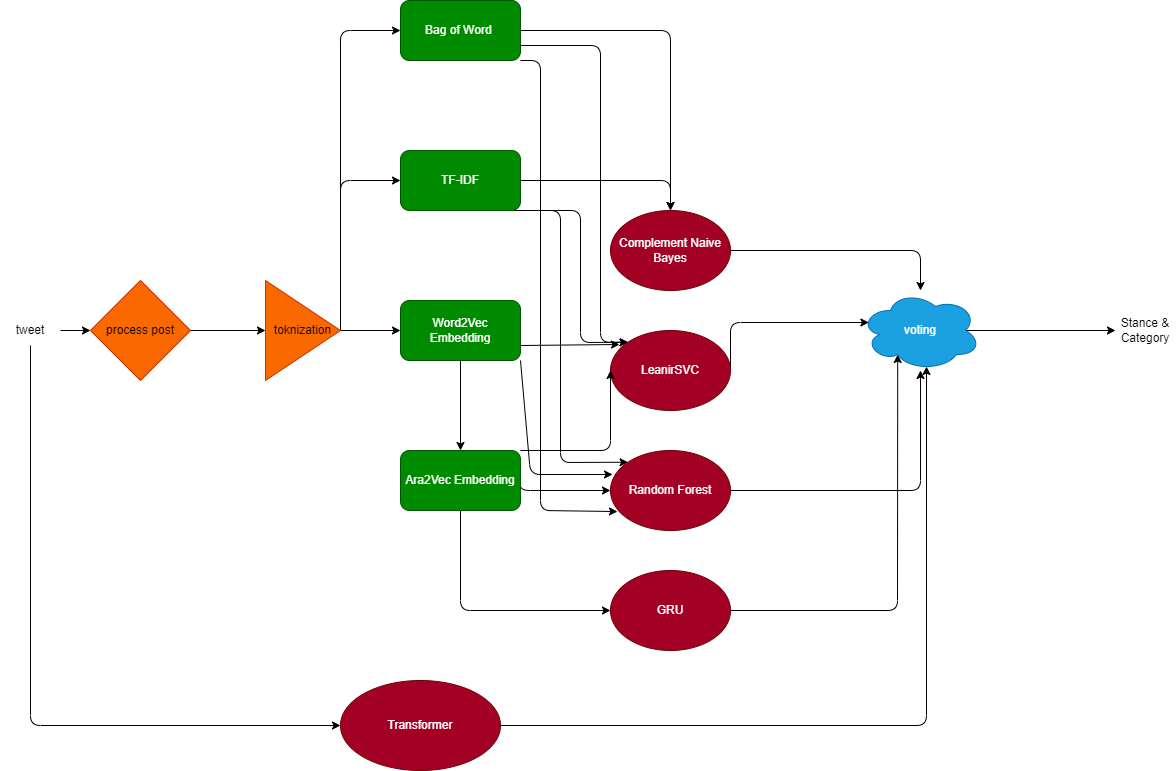
# Project Pipeline:



# A detailed description of each phase in your pipeline

## Data preprocessing:

In data pre-processing we try to remove any unrelated characters; like links, tags (<LF>), stop-words, hash-tags, punctuation, …

Then we used lemmatization to get the roots, we try more than one Lemmatizer but we didn’t find a huge difference

## Feature extraction:

In this phase we try TF-IDF as power full feature especially with classical machine learning algorithm (e.g. Naïve Bayes, SVC, …).

Also we tried Bag of Words and it was less than or similar to TF-IDF

In word embedding, we try to use pre-trained word embedding like Word2Vec then we found pre-trained model on Arabic tweets these embedding help our models to learn faster.

Finally, we use contextual embedding in pre-trained transformer.

## Model training:

Classical models: we tried SVC, Random Forest and Naïve Bayes, and with each model we tried all the features mentioned above.

For Recursive models: we try LSTM and GRU, we try them with one hot encoding first then we use Ara2Vec embedding as an initial hidden layer, finally we choose GRU.

For Deep Learning Recursive models: we used transformers, we try “asafaya/bert-arabic” there was types (mini, medium, …) we choose mini depending on number of parameters and experiments.

# Evaluation: Report the macro F1-score (and all other metrics you tried) for all trials you did.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | TF-IDF | BoW | CBOW | Ara2Vec | Bert-Arabic |
| LinearSVC |  |  |  |  |  |
| Random Forest |  |  |  |  |  |
| Complement NB |  |  |  |  |  |
| LSTM |  |  |  |  |  |
| GRU |  |  |  |  |  |
| Transformers |  |  |  |  |  |

# Specify what model you used for the test set submission on Kaggle and the reason for choosing it.