NLP PROJECT
Arabic tweets

PIPLINE →

ML Architecture

Pipeline

Pre-processing

- Remove Punctuations
- Remove Emoji
- Remove Stop Words
- Remove url
- Remove @username
- lemmatize word

Extract Features

- Bag of Words
- TF-IDF
- Word2Vec Embedding
- Ara2Vec Embeeding

Train Model

- Naive Bayes
- LinearSVC
- Random Forest
- GRU

Text

دعبول حضرتك منو انت وتطلب من قائد دولة إسلامية لقاح لعد ما اتابع الاخبار هم بكل مجالاتهم متفوقين وراح يطلع اللقاح قريباً؟
<LF>#دعبول_دومه_مسحول

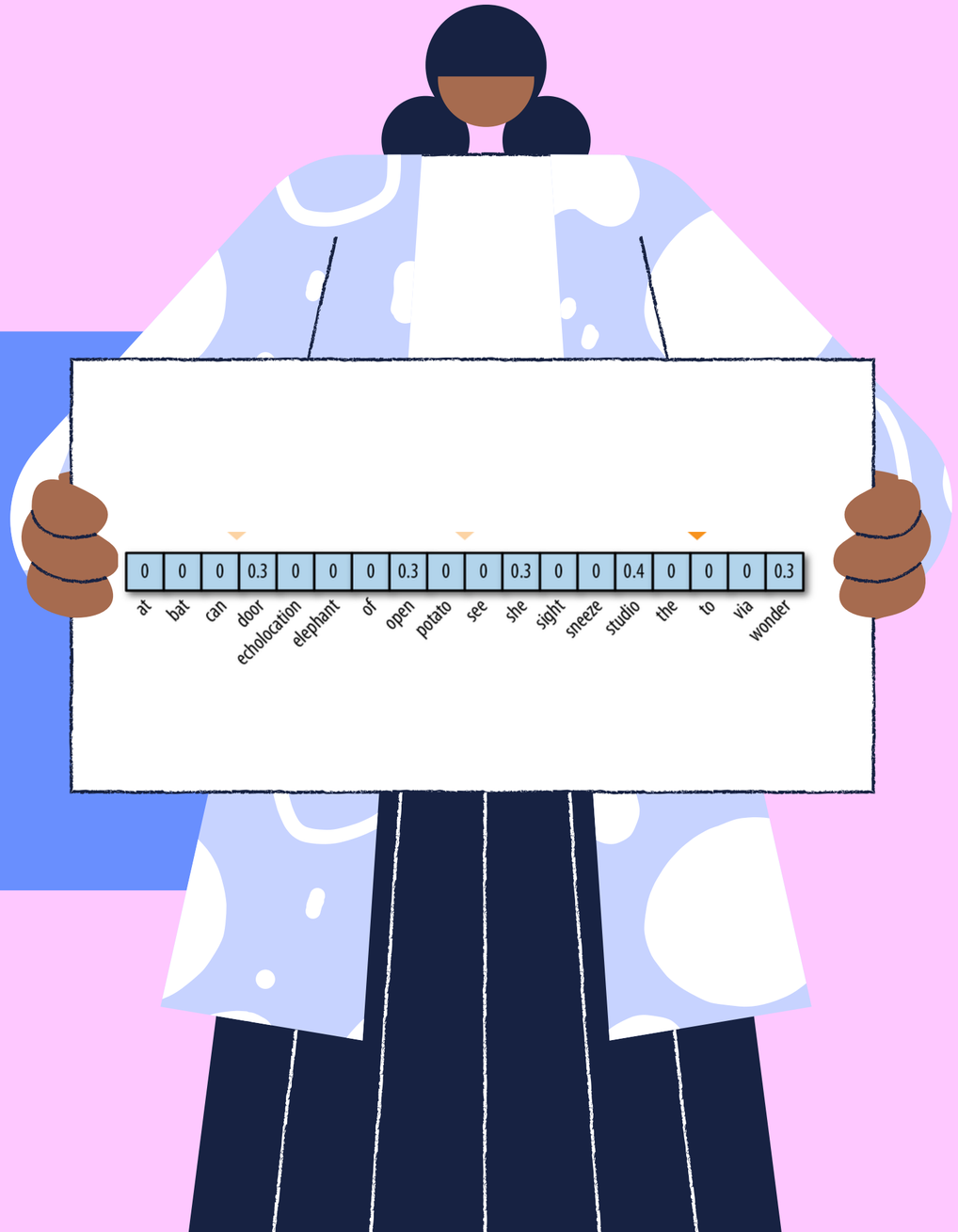
processed text

دعبول حضر من نت طلب قائد دول إسلام قاح
عد تابع اخبار مجال متفوق طلع قاح قريبا
دعبول دوم مسحول

→

Features

- Bag of Words (less accuracy)
- TF-IDF was power full feature in Ml models
- Word embedding (Ara2Vec)



Classical models

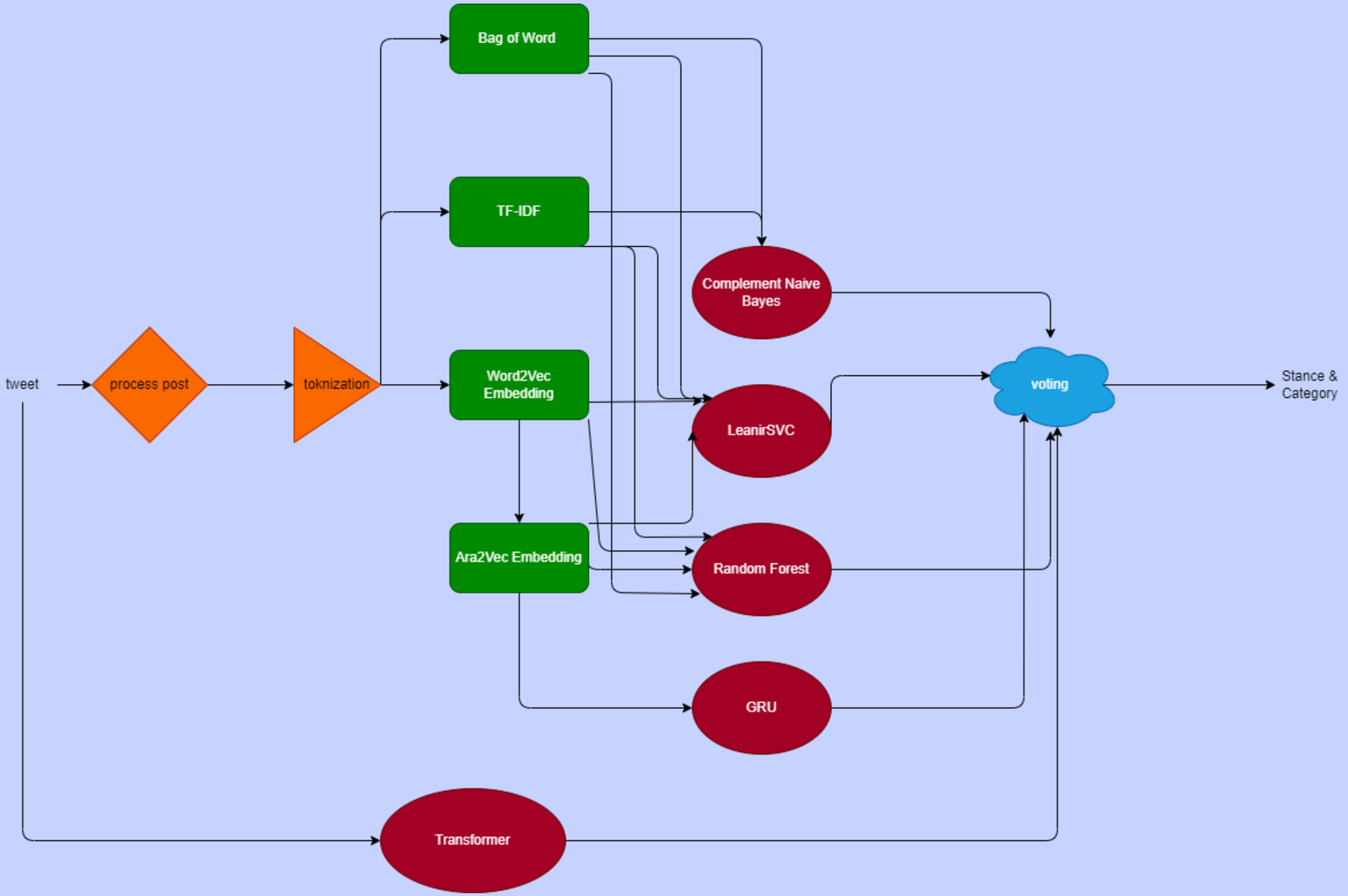
we tried SVC, Random Forest and Naïve Bayes, and with each model we tried all the features mentioned

Recursive models

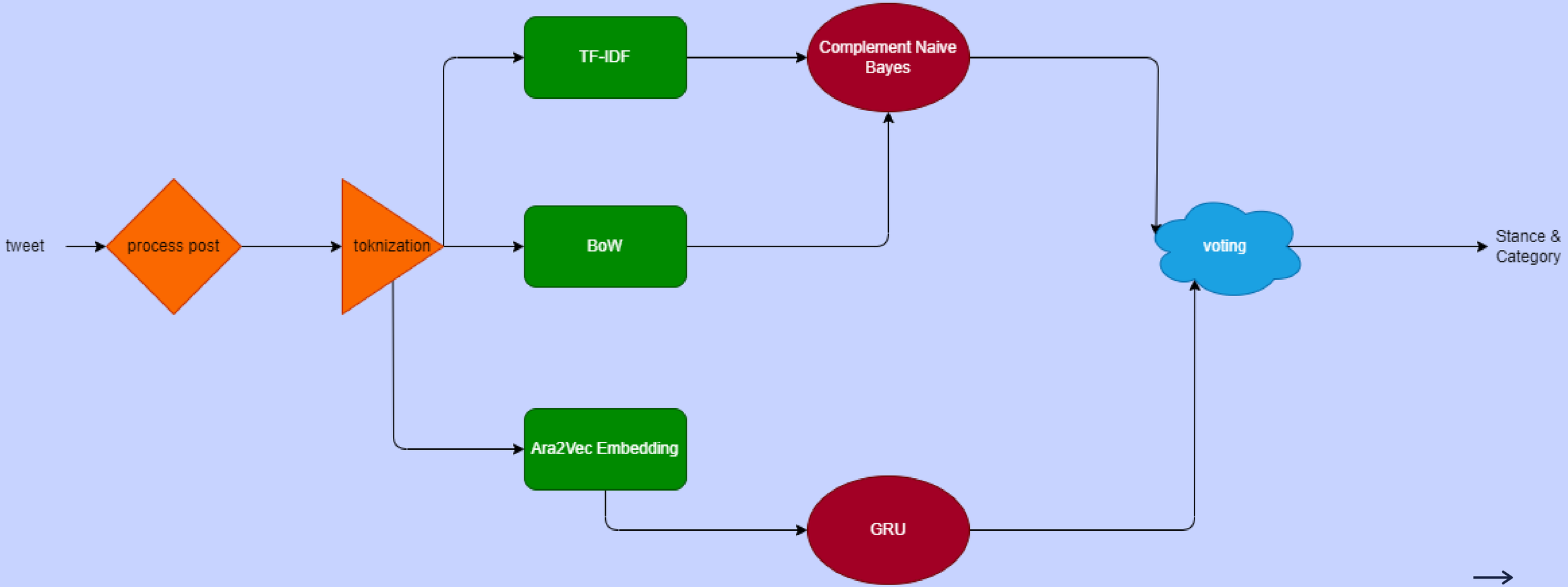
we tried LSTM and GRU, we try them with one hot encoding first then we use Ara2Vec embedding as an initial hidden layer, finally we choose GRU.

→

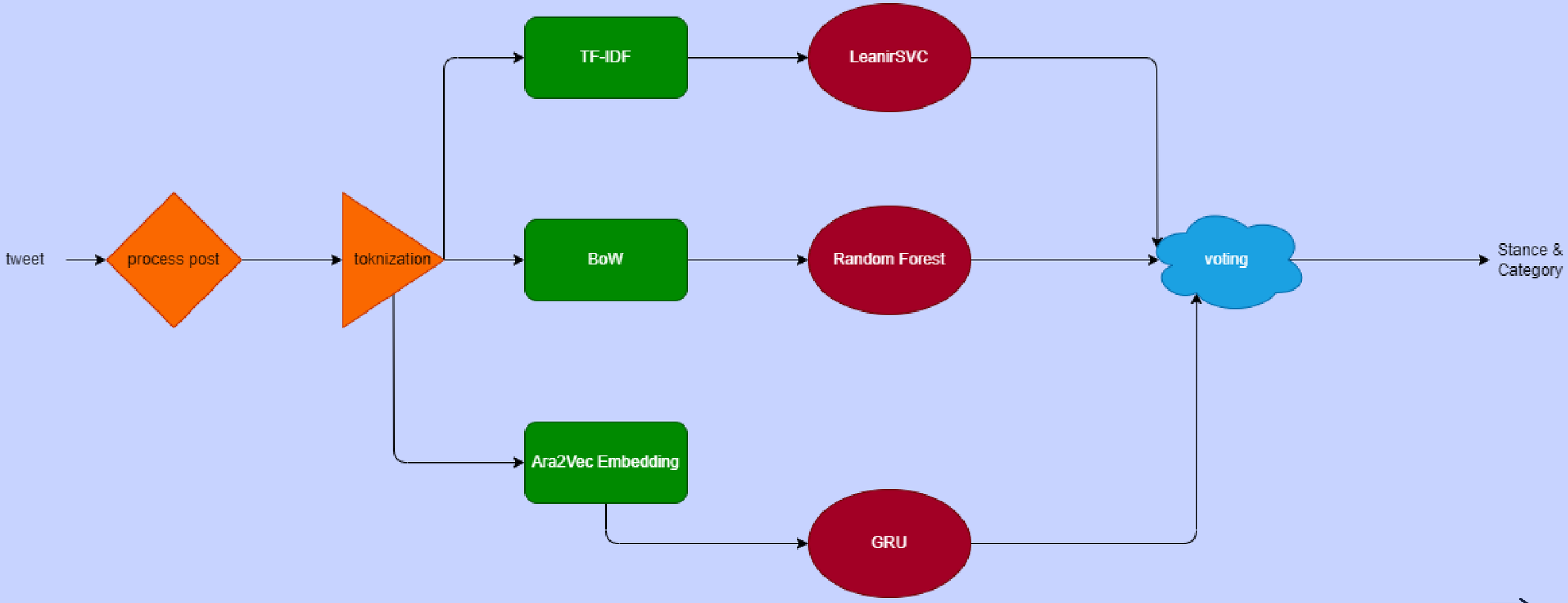
Our Trials


Bag of Word
TF-IDF
Word2Vec Embedding
Ara2Vec Embedding
tweet
process post
toknization
Complement Naive Bayes
LeanirSVC
Random Forest
GRU
Transformer
voting
Stance & Category

Voting Model for Stance



Voting Model for Category



Results On validation

	precision	recall	f1-score	support
-1	0.28	0.31	0.30	70
0	0.39	0.56	0.46	126
1	0.92	0.85	0.88	804
accuracy			0.77	1000
macro avg	0.53	0.57	0.55	1000
weighted avg	0.81	0.77	0.79	1000

accuracy: 0.773

	precision	recall	f1-score	support
0	0.10	0.20	0.13	10
1	0.84	0.84	0.84	145
2	0.76	0.66	0.70	545
3	0.38	0.29	0.33	17
4	0.42	0.63	0.50	128
5	0.40	0.02	0.05	82
6	0.10	0.20	0.13	20
7	0.20	0.50	0.29	2
8	0.00	0.00	0.00	15
9	0.23	0.61	0.33	36
accuracy			0.60	1000
macro avg	0.34	0.40	0.33	1000
weighted avg	0.64	0.60	0.60	1000

accuracy: 0.598

→

DL Model

- Encode the lables
- used (pretraines) Tokenizer to tokenize the text and get inputs
- used DataLoader to load train and validation data
- tune the model parameters in 3 epochs (11.6 M Total params)
- learning rate =.0001
- maxlen for tweet =80

arabic-bert-mini model was pretrained on ~8.2 Billion words

Results On validation

	precision	recall	f1-score	support
-1	0.39	0.44	0.42	70
0	0.53	0.55	0.54	126
1	0.92	0.90	0.91	804
accuracy			0.82	1000
macro avg	0.61	0.63	0.62	1000
weighted avg	0.83	0.82	0.83	1000

	precision	recall	f1-score	support
advice	0.14	0.50	0.22	10
celebrity	0.80	0.88	0.84	145
info_news	0.74	0.66	0.70	545
others	0.00	0.00	0.00	17
personal	0.53	0.58	0.55	128
plan	0.23	0.32	0.27	82
requests	0.50	0.20	0.29	20
restrictions	0.00	0.00	0.00	2
rumors	0.12	0.13	0.12	15
unrelated	0.33	0.42	0.37	36
accuracy			0.61	1000
macro avg	0.34	0.37	0.34	1000
weighted avg	0.63	0.61	0.62	1000

Thank You

Team Members

Donia Abd El-Fatah
Raghad Khaled
Menna Allah Ahmed
Nada El-Sayed