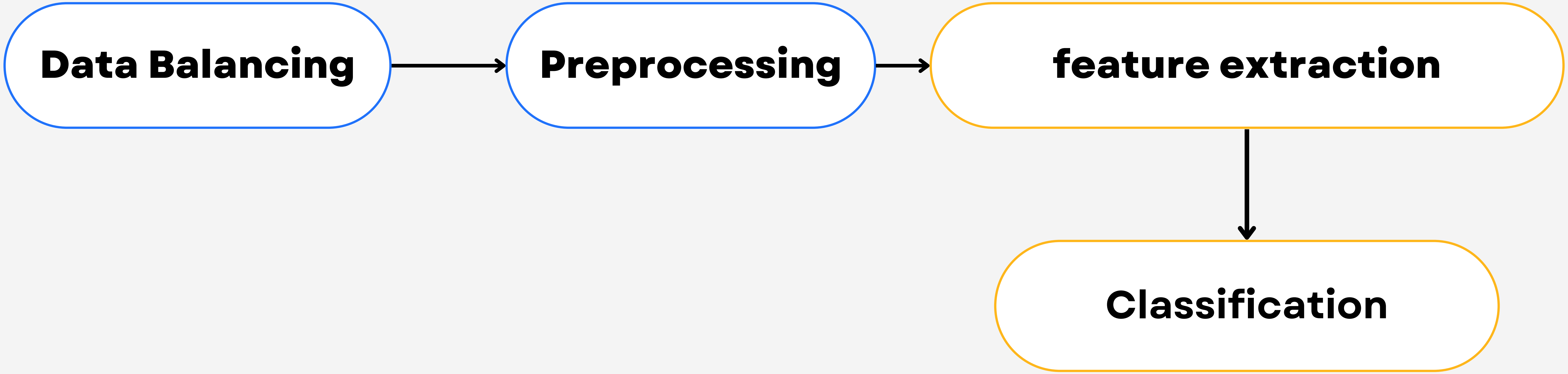
Arabic Tweets Stance Detection and Category Classification

Team number: 12

Presented to:
Eng. Omar Samir

Project Pipeline



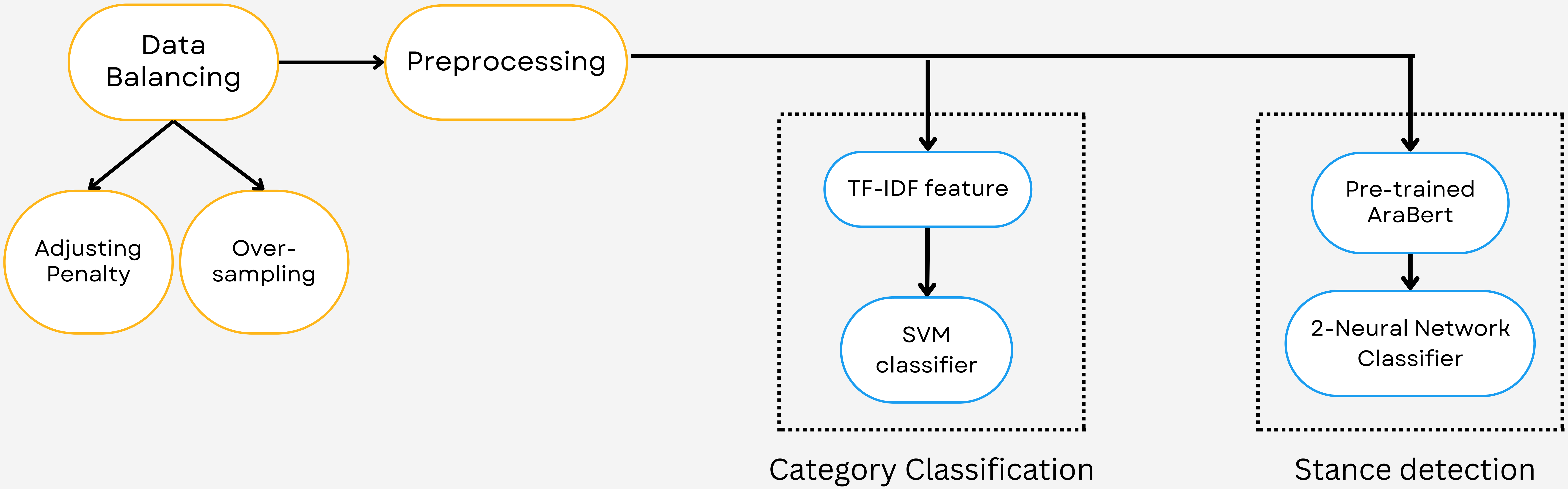
Data Balancing

Preprocessing

feature extraction

Classification

Project Pipeline



Data Balancing

Adjusting Penalty

Over-sampling

Preprocessing

TF-IDF feature

SVM classifier

Category Classification

Pre-trained AraBert

2-Neural Network Classifier

Stance detection

Analyzing Data

Stance classes is very unbalanced!

Class	percentage	support
positive	0.792501	5538
neutral	0.144820	1012
negative	0.062679	438

Most of the data are positive tweets.

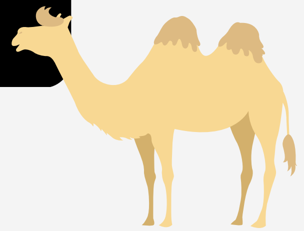
ض

Category classes is very unbalanced!

class	percentage	support
info_news	0.517459	3616
personal	0.146680	1025
celebrity	0.139525	975
plan	0.086720	606
unrelated	0.046222	323
others	0.023898	167
requests	0.016027	112
rumors	0.011305	79
advice	0.009588	67
restrictions	0.002576	18

م

ح



Data Balancing

We've followed 2 approaches to handle the unbalanced dataset:

- **Oversampling**
  - More samples for minoriy classes
- **Penalizing mistakes**
  - Higher penalty for minority classses

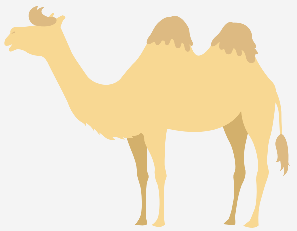
Problem was un-avoidable as unlike 'accuracy', 'macro f1' score will just collapse when we ignore some very low probabilty classes,

م ض ح

Data Preprocessing

- Removing diacritization, punctuation and normalizing letters
- Replacing links, numbers and mentions with <link>, <num> and <mt>
- Converting emojis to equivalent text (😂 -> face_tearing_with_joy)
- Lemmatization
- Removing stopwords e.g. 'معي' , 'عندنا' , 'وأيها'
- Converting English text to lowercase
- Emphasizing Hashtags

م ض ح



Feature Extraction

- Bag of words (BOW)
- Continuous BOW (word embeddings / vectors)
- Skip-gram (word embeddings / vectors)
- TF-IDF
- Arabert Embeddings as a feature for SVM

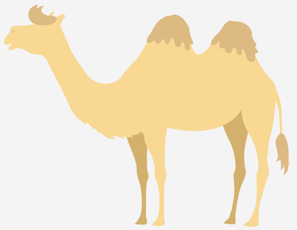
Models

- Classical Models
- Sequence Model
- Transformers

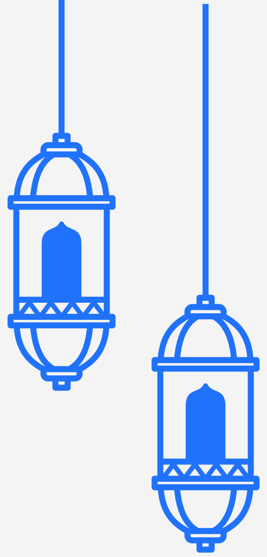
م

ض

ح

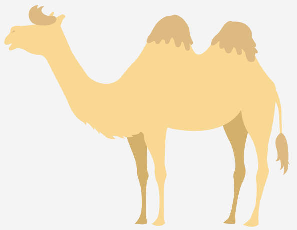


Classical Models



- SVM
- Naive Bayes
- KNN
- Decision Trees
- Random Forest   n_estimators = 1000
- Logistic Regression n_iterations = 300

م   ض   ح



Classical Models Trials

Data	Features	Classifier model	Acc	F1	S/C
Original Data	BOW	Random Forest n_estim=1000	80	46	S
Original Data	BOW TFIDF_W TFIDF_C	Logistic Regression balanced	77	54	S
Original Data	CBOW	Naive Bayes	51	31	S
Original Data	SG	KNN k=5	71	36	S

م ض ح

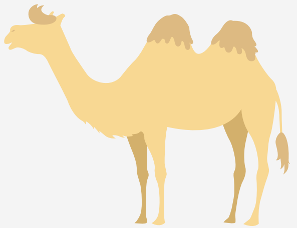
Sequence Models

Approach

- An Embedding layer
- 3-layer LSTM
- 1 linear neural network layer for classification.

Accuracy	F1-score	Problem
54.2	27.2	Category Classification

م ض ح



Transformers

Fine-tuned an arabic bert model on our dataset.

Model name: aubmindlab/bert-base-arabertv02-twitter
Pretrained on **~60 Million Arabic tweets**.

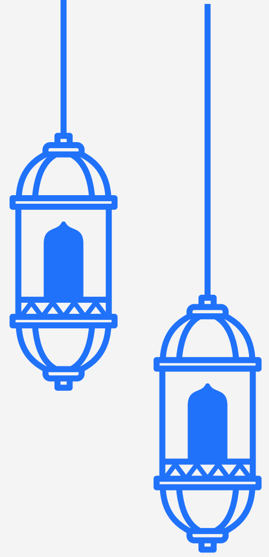
Approach

- Freezing the bert’s parameters
- Produce the embedding as bert's pooled_output.
- Classifier head that consists of 2 neural network layers.

Accuracy	F1-score	Problem
77.6	61	Stance Detection

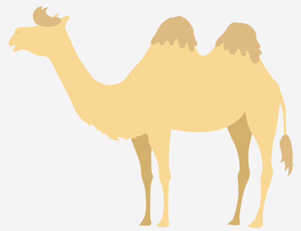
م ض ح

Results



Data	Features	Classifier model	Acc	F1	S/C
Original Data	**HF - Arabert with lower learning rate**		**84.1**	**65.2**	Stance
OverSampled	**TFIDF_C TFIDF_W**	**Linear SVM Balanced, farasa lemmatize + non-lemmas**	**60**	**34**	Category

م ض ح



م

شكراً

ض

ح