Machine Learning Engineer Nanodegree

Capstone Project**: Building a sentimental analysis model for Arabic dataset**

Rabab Alkhalifa – raalkhalifa@iau.edu.sa

I. Definition

Sentiment analysis is a sub-field of natural language processing and employs machine learning, computational linguistics and data mining. The goal of sentiment analysis is to automatically detect the polarity of a text and try to systematically identify, extract, quantify, and study affective states and subjective information within text.

II. Project Overview

In research will build upon existed Arabic dataset from the web which were designed for detecting emotions for text, the goal is to optimize existing models toward designing a robust computational approach for analyzing and detecting emotions from Arabic dataset. Additionally, conducting research on the state of art techniques for Arabic sentimental analysis. The outcome from this research will contribute knowledge towards recognize, understand and detect accurate emotion from Arabic dataset called Arabic Sentiment Tweets Dataset (ASTD).

III. Problem Statement

This research project aims to explore the most accurate model to classify Arabic sentimental dataset. This will be look at through the evaluation for different proposed models to contribute in designing a robust model for evaluating emotions.

IV. Metrics

We described the performance measurements by using numerical scalars and ROC curve to easily describe the accuracy, F-score and contingency table of the selected models.

a. **Accuracy** is good measurement if test sets fully describe the diversity of data tuples to be classified.
b. Precision and recall ignore true negatives, which focus more in measuring the performance. We will use the **F1 score** which can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst at 0.
c. **Contingency table** will be constructed in the aim of measuring true positive, turn negative, false positive and false negative. Then, precision, recall and F-measure accurately analyzed to avoid selecting model that always predict negative.

V. Analysis

a. Data Exploration

The dataset has been explored first by converting it to CSV file and read it through DataFrame. The dataset consists of 10006 tweets, each categorized as OBJ, NEG, POS, NETURAL (Table 1). The are no empty data nor inconsistency between samples' classifications. While the prevalent category is OBJ with about 6675 tweets.

Table 1 Sample from ASDA Dataset

	Tweet	Classification
0	...بعد استقالة رئيس #المحكمة_الدستورية ننتظر استق	OBJ
1	...أهنئ الدكتور أحمد جمال الدين، القيادي بحزب مصر	POS
2	...البرادعي يستقوى بامريكا مرةاخرى و يرسل عصام ال	NEG
3	... الحرية_والعدالة | شاهد الآن: #ليلة_الاتحادية#	OBJ
4	...الوالده لو اقولها بخاطري حشيشة تضحك بس من اقول	NEUTRAL

To do more exploration, we investigate the existence of stop words in our dataset and found out that there is no need to remove any because there is no match between Arabic stop words in NLTK package and our dataset as such we skipped this step (code is commented in the notebook) and the fact that we were using the cleaned data from ASDA dataset. Also, to get better insight, frequency of each word calculated using CountVectorizer and 1-gram this allow us to construct a DataFrame consist of word-frequency for all the existent words in the dataset (Figure 1).

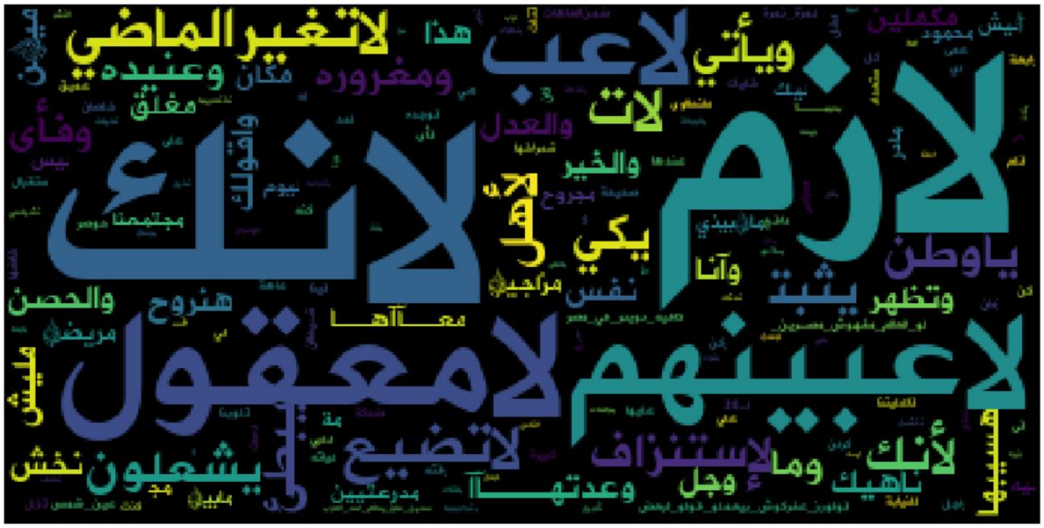
Figure 1Word-Frequency of ASDA dataset

	frequency
دراما	17941
استنتاجات	3519
الغربة	8174
إفساد	2405
شيفنا	20447

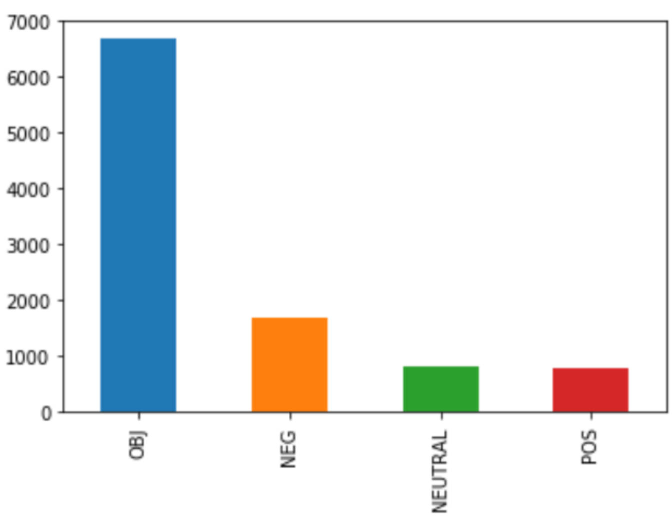
b. Exploratory Visualization

In order to explore the data, we used word cloud with word frequency to generate the most repeated words (Figure 2).

Figure 2 ASDA Dataset word frequency



The data distribution in in balanced classes is visualized using bar chart, OBJ is 6675, NEG is 1682, NETURAL is 831 and POS is 799.



With Down Sampled data (see VII. Data Preprocessing) we down sampled our data to 799 to match minority class POS. While with up sampled data we up sampled them to match the majority class OBJ to 6675.

c. Algorithms and Techniques

In our experiments, we used ten standard classifiers, eight of them were already specified ASDA research paper (read d. Benchmark section). The other two are Random Forest and Decision Trees as we wanted to evaluate their performance compared to the others.

d. Benchmark

In this research, we will use ASDA designed and proposed in [1] as the benchmark model. ASDA stands for Arabic Sentimental Tweet Dataset which is used for Arabic social sentimental analysis. The dataset Consist of more than 10000 tweets gathered from twitter. Each tweet is Classified as: objective, subjective positive, subjective negative or subjective mixed.

The benchmark model was built and evaluated using corpus collected from twitter using eight standard supervised classification models with balanced classes (Figure 3 and Table 2) and unbalanced classes (Figure 3 and Table 3).

Figure 3 Balanced Data Distribution using Bar Chart

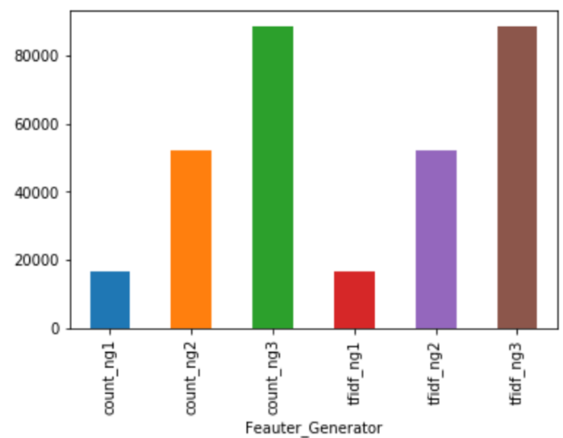


Table 2 Balanced Data feature count for each generator

DataSet Name	Feauter_Generator	Feauters_Count
4-balanced	count_ng1	16457
4-balanced	count_ng2	52036
4-balanced	count_ng3	88672
4-balanced	tfidf_ng1	16457
4-balanced	tfidf_ng2	52036
4-balanced	tfidf_ng3	88672

Figure 4 Un-Balanced Data Distribution using Bar Chart

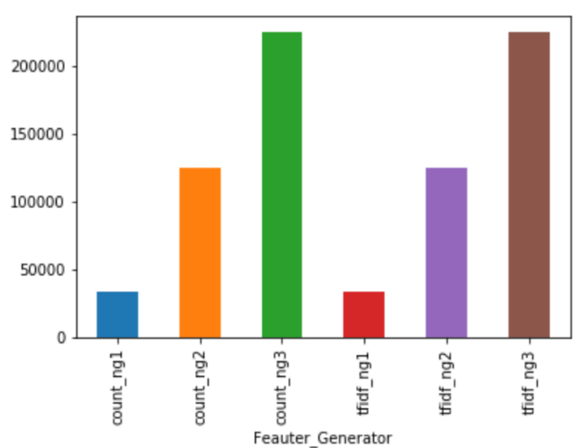


Table 3 Un-Balanced Data feature count for each generator

DataSet Name	Feauter_Generator	Feauters_Count
4-unbalanced	count_ng1	33365
4-unbalanced	count_ng2	124760
4-unbalanced	count_ng3	225111
4-unbalanced	tfidf_ng1	33365
4-unbalanced	tfidf_ng2	124760
4-unbalanced	tfidf_ng3	225111

The Classifiers used in their experiments are: Logistic Regression, Passive Aggressive, Linear Support Vector Machine, Perceptron, BNB, MNB, SGD, K-Nearest Neighbors. All of them used with the dataset taking 20% for testing without parameter tuning.

The dataset in their experiment is preprocessed. They cleaned text by removing special characters, non-Arabic text, white spaces and stop words. They used token count and Frequency–

inverse document frequency (Tfidf) explained in IX. Data Preprocessing section to generate features.

VI. Methodology

In order to achieve a solution for the problem statement, the main objectives we followed a clear plan. First, Review the literature to evaluate various experiments and datasets that was used to extract sentiments from Arabic datasets and discuss the outcomes of experimentation by replicating selected experiments model with ASTD dataset. Then, create an enhanced sentimental analysis model that is an upgraded version from the existing models that proved to gain better insights from sentimental dataset. This include a model to compare a set of reviewed experimented models and enhanced model developed after observation. This is done through tuning the parameters using Grid Search method in python and applied in both the standard architecture (replicated experiment) and the enhanced model (using different preprocessing steps and enhancing and testing the standard models before and after parameters tuning). Finally, we statistically evaluate, analyze and interpret the results obtained to assess the accuracy of sentimental analysis models using Arabic dataset. In the below sub-sections the technical detailed implemented in our experiment.

a. Data Preprocessing

In order to processes the data, we implement three different procedures, String token count (Bag of Word), Frequency–inverse document frequency (Tfidf) and data sampling.

First, we implement string token count using Sklearn library, CountVectorizer. It takes each sentence (all the words) present in the data set in the review section and then splits each of the words present in the form of tokens. The occurrence of these tokens in the whole data set are counted in such a way that the count of the occurrence of each token in a positive, negative, natural or objective feedback. Then they are collected separately. Finally, the word frequency of the tokens is calculated.

Then, after splitting the data we calculate Frequency–inverse document frequency (Tfidf). Datas et predicted without Tfidf called count_ng1, count_ng2 and count_ng3, while dataset predicted using Tfidf called tfidf_ng1, tfidf_ng2 and tfidf_ng3. The numbers refer to the features range. If n-gram equals to one, then each unique word will be one feature, while ng3 means that one wo rd, two words and three words each as a bag will be considered three different features. Thus, t he number of features increases with the increase of n-grams (more explanation in b. benchmar k). One advantage of Tfidf is that it measures how important a word is to differentiate each cate gory. It reduces the weightage of more common words like (stop words or common words) whic h occurs in all tweets.

Finally, to sample from our data and to follow ASDA research paper, the data was down sampled and up sampled to generate a balanced dataset, in addition to using the unbalanced dataset (statistics described in Exploratory Visualization).

In up sampling, we up-sampled the minority ratings these are POS, NEG and NEUTRAL. First, we separate samples from each class into different DataFrames. Next, we resampled every minority class with replacement, setting the number of samples to match that of the majority class. Finally, we combined the up-sampled minority classes' DataFrames with the original majority class DataFrame. The process is like that of up-sampling, but we down sample all the majority classes to the minority class POS, without replacement as there is no need for it when data is reduced.

b. Implementation

we started by replicating the experiment in ASDA research paper. Then, we add two more classifiers to have a total of ten different classifiers. Additionally, we used the hyper tuning parameters to give our classifier different options and increase the chance of getting better accuracy score. This all done by starting with the github code of ASDA but we changed it to match our environment settings and moved part of it in jupyter notebook. The outcome from this step is two jupyter notebooks. First one, we named it *sentimental_analysis_BenchMarkModel_ASTD*. It is based on the original parametric settings which give accuracy and F-score similar to the numbers mentioned in the research paper as it used the same dataset splits and parametric settings. The second one, we named it *sentimental_analysis_Model_ASTD_BestResults*. It is based in our data preprocessing steps mentioned above and ASDA dataset cleaned reviews. It examines model performance using all the ten classifiers tuned using GridSearch.

c. Refinement

Our method in up and down resampling dataset out performed the balanced data in classification accuracy reported in ASDA research paper. The most obvious differences between our model and ASDA model is that they used standard classifiers without cross validation nor parameter tuning, in addition to dis-similarity in resampling methods. For our experiment, we choose to up sample the data and down sample it to make it balanced (check Data Preprocessing section). In ASDA research paper, they did not explicitly state their methodology for balancing the data, however, it seems to us that it was manually manipulated.

VII. Results

a. Model Evaluation and Validation

Our model examined the accuracy using cross validation and can predict the data with accuracy reached 70% higher than the accuracy reported in ASDA research paper using un-balanced data and LinearSVC with two folds. Our concerns that even through accuracy is getting higher the F-Score measurement is not (Figure 5).

Figure 5 Down-sampled and Un-Balanced ASDA dataset

Also, the model reaches all benchmark model with the down streamed data with accuracy as high as %56 with BNB (Table 4).

Table 4 Down Balanced ASDA model performance before and after parameter tuning

	clasifiers	data	feauter_generator	UnOpt_accuracys	UnOpt_fScores	Opt_accuracys	Opt_fScores
4	bnb	down_balanced	count_ng1	0.567187	0.563	0.567187	0.563569
14	bnb	down_balanced	count_ng2	0.567187	0.568337	0.567187	0.568339
34	bnb	down_balanced	tfidf_ng1	0.567187	0.563	0.567187	0.563569
44	bnb	down_balanced	tfidf_ng2	0.567187	0.568337	0.567187	0.568339

Upstream Sampling out preform all the above with accuracy as high as %97 with RandomForest using 3-ngram range of features (Table 5).

Table 5 Up Balanced ASDA dataset performance

	clasifiers	feauter_generator	UnOpt_accuracys	UnOpt_fScores	Opt_accuracys	Opt_fScores
7	RandomForest	count_ng1	0.960209	0.959525	0.97422	0.973823
12	SVM	count_ng2	0.969923	0.969272	0.970297	0.969717
17	RandomForest	count_ng2	0.96488	0.964357	0.975528	0.975166
20	Logistic Regression	count_ng3	0.971044	0.970465	0.973286	0.972798
21	Passive Aggresive	count_ng3	0.970297	0.96985	0.970484	0.970042
22	SVM	count_ng3	0.97366	0.973239	0.974594	0.974148
27	RandomForest	count_ng3	0.964506	0.964174	0.978517	0.978283

b. Justification

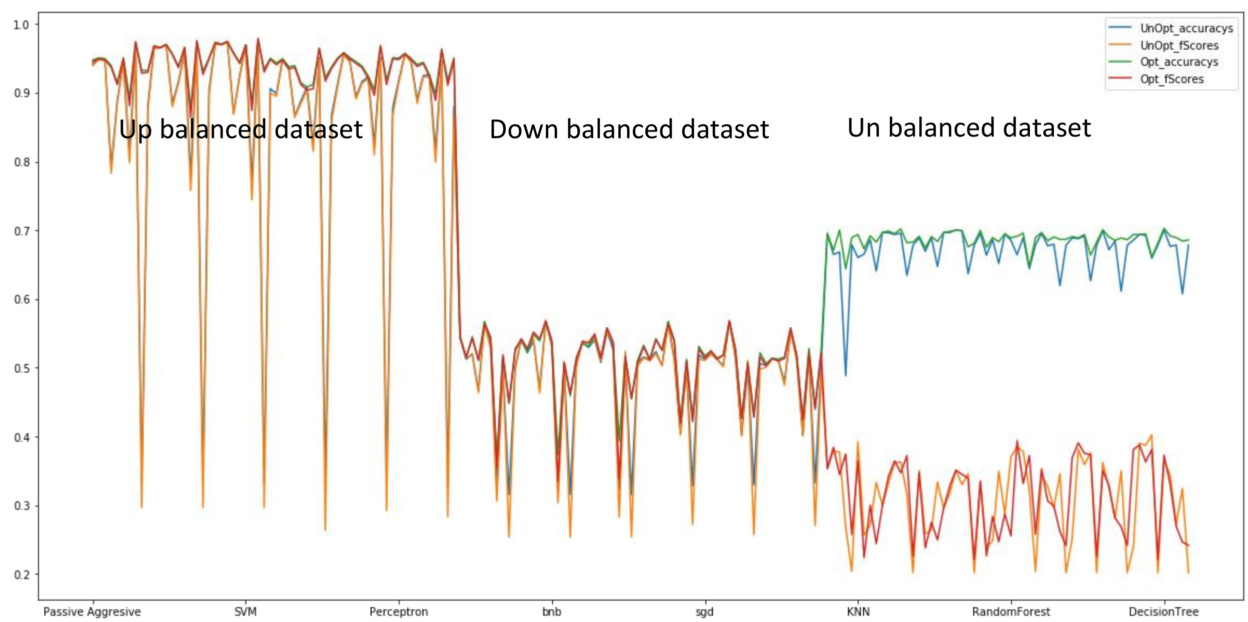
We clearly can see the up-balanced sampled dataset out perform all our models using the ten classifiers, despite of using Grid Search to tune the hyperparameters. However, it highly likely that this model will suffer from overfitting. This due to the fact we increased the number of copies to reach the majority class OBJ for all the other target classes.

VIII. Conclusion

a. Free-Form Visualization

In our experiment we succeeded in improving the performance of the benchmark model which reaches as high as %69.4. Our model using un balanced ASDA dataset reached %70, with significant drop of recall and precision with F-Score merely reached %38. Using down balanced dataset our model successfully competed with the benchmark with accuracy and F-score higher than %49.3 reported in ASDA research paper and reached around %56.7 for both measurements. The highest classification accuracy and F-Score gained with up balanced data with accuracy and F-score above %95. The shows the performance of our ten classifiers using up balanced, down balanced and unbalanced ASDA dataset respective.

Figure 6 Performance of our ten classifiers with up-balanced, down-balanced and un-balanced ASDA dataset.



b. Reflection

In our experiment, we started by replicating ASDA research paper to get the same results as they reported. Once we finalized their work, we started our own by implementing the steps mentioned in the Data Preprocessing section. We find out that Up Sampling increases the accuracy while the model will be vulnerable for overfitting for unseen data.

c. Improvements

We aim to improve our model by measuring the time each classifier takes. We better reduce the number of features using feature reduction techniques and may use Sparks to increase performance.

Additionally, we need to increase the dataset and test it using unseen corpus to get better insight about its performance, especially with the classifiers trained using the up balanced dataset.

IX. References

[1] Nabil, M., Aly, M., & Atiya, A. (2015). Astd: Arabic sentiment tweets dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 2515-2519).