



Faculty of Engineering & Technology – Electrical & Computer  
Engineering Department

First Semester 2022 – 2023

Artificial Intelligence – ENCS3340

*Tweet Emotion Detection Project (Machine Learning)*

---

Prepared By:

Ali Mohammad – 1190502

Ahmaide AlAwawdah – 1190823

Instructor:

Mr. Aziz Qaroush

Section: I

Date: February 2023

## **Abstract**

The main objective of this project is to classify Arabic tweets to be either positive or negative tweets, by analyzing their featured emotions in order to feed and train the system with its needed data to make it a predicting system with high classification accuracy, which will also be tested.

The used programming language for this project is Python, using version 3.9, with the use of some useful libraries such as nltk, ar\_correction, pandas, sklearn, and numpy.

## ❖ Table of Content

1.Problem Specification .....	1
2.Project Stages.....	2
2.1. Data Gathering .....	2
2.2. Data Preprocessing .....	2
2.3. Features Extraction .....	5
2.4. Training & Testing .....	6
2.5. Testing Calculations .....	7
2.6. The Deployment.....	7
3.Running The Program .....	8
3.1. Data Preprocessing Results .....	8
3.2. Feature Extraction Results .....	8
3.3. Classifications Models Results .....	13
4.Conclusion .....	20
5.References .....	21

## ❖ Table of Figures

Figure 1-1: Tweet Classification .....	1
Figure 2-1: Emotion Detection System Diagram.....	2
Figure 2-2: Data Preprocessing .....	3
Figure 2-3: Stemming Process <sup>[2]</sup> .....	4
Figure 2-4: Feature Extraction Diagram <sup>[3]</sup> .....	5
Figure 2-5: Training and Testing Diagram .....	6
Figure 3-1: Data Processing Outcome.....	8
Figure 3-2: Emoji Counting Difference .....	9
Figure 3-3: Hashtag Counting Difference .....	9
Figure 3-4: Sentence Counting Difference .....	10
Figure 3-5: Character Counting Difference .....	10
Figure 3-6: Word Counting Difference .....	10
Figure 3-7: Hashtag Percentage Difference .....	11
Figure 3-8: Emoji Percentage Difference .....	11
Figure 3-9: Average words in a sentence Difference .....	11
Figure 3-10: Average Characters in a Word Difference .....	12
Figure 3-11: Stop Words Counting Difference .....	12
Figure 3-12: Random Forest Confusion Matrix.....	13
Figure 3-13: Random Forest True Positive Graph .....	13
Figure 3-14: Random Forest False Negative Graph .....	14
Figure 3-15: Random Forest Precision-Recall Graph .....	14

## 1. Problem Specification

In this project the problem is specified to be the classification of Arabic tweets, where a tweet can be classified as either positive or negative tweet judging by its content (text and emojis).

The content should display the needed information that can describe the emotions in this tweet in order for the system to detect it and decide the situation of the tweet (positive or negative).



Figure 1-1: Tweet Classification

First the tweets content is full of unnecessary data that may damage the tweet's classification, and for that the tweets will need some processing in order to get rid of these unnecessary contents to give the tweet the ability to contribute in training the system or to be able to be classified.

In order to determine whether if a tweets content is either positive or negative, a set of features must be defined, where these features depend on the tweet's emojis and text content as the system measures the statistics of these features for a good amount of both positive and negative training tweets, so when they're measured for a given tweet the system shall decide if its positive or negative.

So, in order to make a complete classification system, two sets of data (tweets) will be needed with each tweet being classified as if it is actually positive or negative, where the first set will train the system to classify the tweets and the second set will test the work flow of the system after it was trained with the first set.

## 2. Project Stages

For this project as shown in figure 2-1, the system is divided into multiple parts where each part plays a part that takes the system to its needed outcomes.

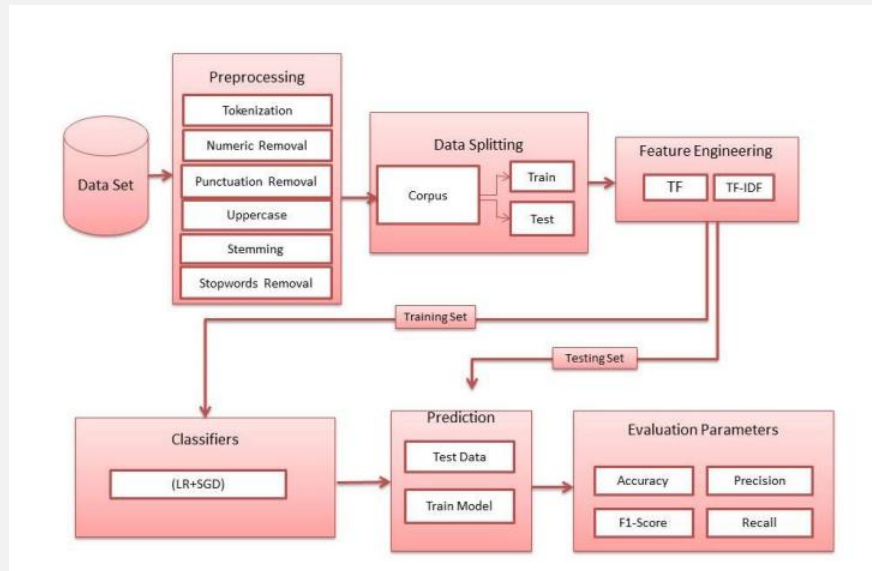


Figure 2-1: Emotion Detection System Diagram

### 2.1. Data Gathering

In order to build a system that detects the emotions in Arabic tweets tweet it, first it needs to be fed by given gathered tweets which with their classifications being given too, in order to train the system with a percentage of tweets for each class, and then tested with the remaining tweets.

In this project there two data files are inserted where one of them will contain the gathered positive tweets, and the other will have the negative tweets.

### 2.2. Data Preprocessing

After gathering the data, before any of the training or testing processes, that data will include many components that needs to be cleared out in order to have clean data that can be processed so it can have better clarifying of its features, and for that the data will need to be preprocessed (cleaned) using the nltk library in python, some features are calculated before the text being fully processed (will be detailed) , figure 2-2 below shows an example of an Arabic text preprocessing.

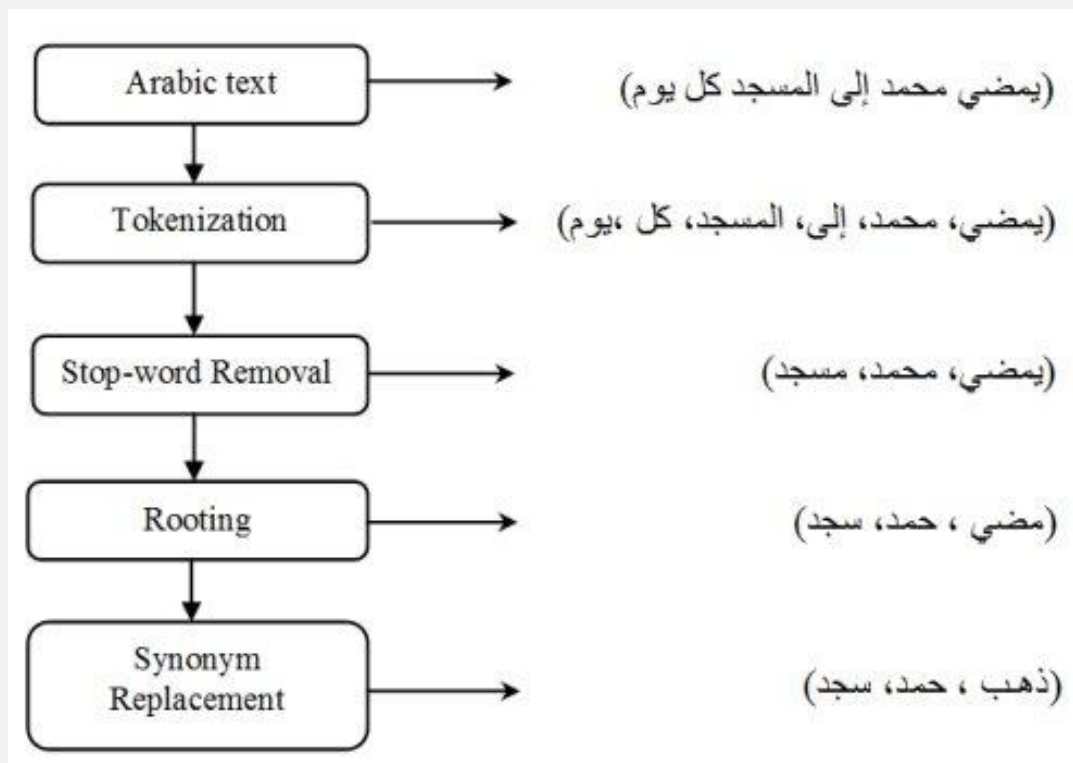


Figure 2-2: Data Preprocessing

Here are the operations that are done to the data in order for it to be preprocessed:

### 2.2.1. Emojis Extraction

As said before these Arabic tweets can contain emojis that can be useful in the tweet's emotion detection, emojis such as:



These emojis will be extracted from the text in order to decide their features separately.

### 2.2.2. Removing Hashtags

Hashtags such as ( #ديوان\_الشعر، #مساء\_الخير\_للجميع، #الاتحاد ) won't help in the emotion detection process yet it might damage it, which makes them useless, so they will be extracted from the text, yet before the removal the number of hashtags in each tweet is calculated in order to be used as a feature.

### 2.2.3. Removing Characters other Than Arabic Letters

Any character in the tweet's content such as other language character (English, Spanish, etc.) or any numeric character is removed from the text, as it won't give any detail about the emotions in the text.

#### 2.2.4. Remove Vocalization

As known Arabic words can have some vocalization on its letters such as (بَ، بٍ، بٌ، بُ، ) (بِ، بٍ، بٌ) these vocalizations are not useful in coding as they will make the strings characters more complicated so they will be removed.[1]

#### 2.2.5. Give each Arabic letter one Shape

As known in the Arabic letters, a letter can have more than a state or shape such as "ا" can be any of the following (أ، إ، آ، ا) so these kinds of letters should be formalized to have only one shape or one state.

#### 2.2.6. Removing Consecutive Characters

Consecutive characters in the tweet's texts are also remove.

#### 2.2.7. Removing Stop words

Stop words are also a part of blocking the emotion detection process so the stop words are removed from the text after setting up that the stop words that are being searched for are Arabic stop words, stop words such as (يا، أكثر، الذي، من، على، في، إن، أي، بعد، لولا، نحن، يا).

#### 2.2.8. Tokenization

The tokenization process is simply separating the text words from each other, as it turns the string into a list of its content, which will make the feature extraction process easier.

#### 2.2.9. Stemming

Stemming is a process that gives the original meaning of each word with removing its extra, this can help in extracting the similarities between the same classification tweets, as shown in figure 2-3.

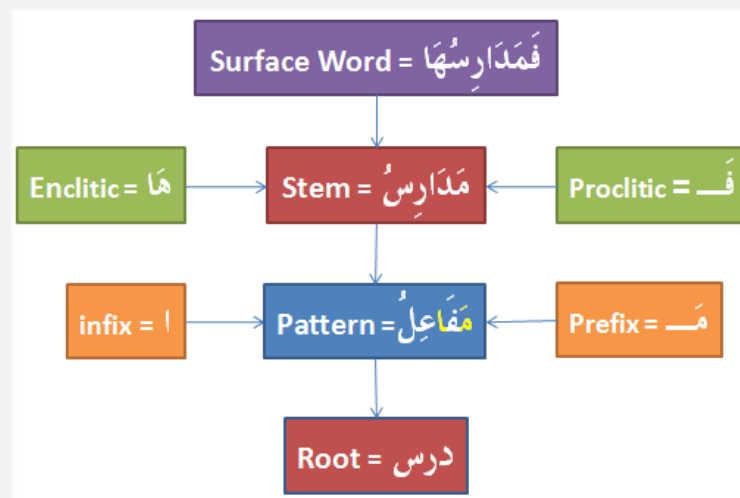


Figure 2-3: Stemming Process [2]



## 2.3. Features Extraction

Each tweet has its own contained features, these features can play a big role in detecting the tweets emotions, in order for it to be classified, features can depend on both the tweet's text content, and the content emojis too, as each one of them can produce its own features.

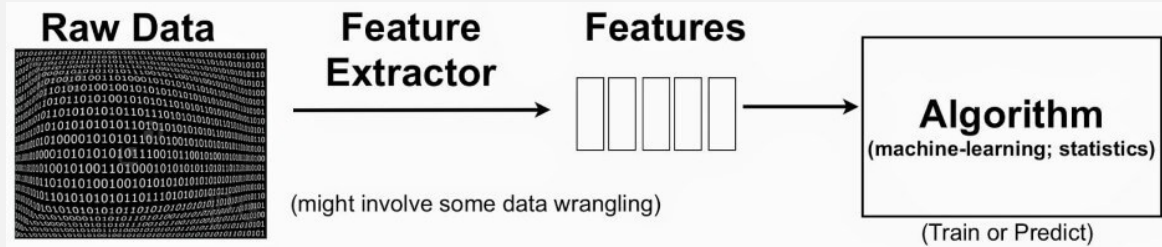


Figure 2-4: Feature Extraction Diagram [3]

### 2.3.1. Manually Extracted Features

Some features were extracted manually in order to increase the performance of the models, since the automated extractions only measures the frequency of the words, the following features were included in this machine learning system manually:

- 1. Emoji Count:** The number of the included emojis in a tweet can be a feature that is used in
- 2. Hashtag Count:** Same as the emojis the number of the included hashtags in the tweet can also be a feature that is used in classification.
- 3. Sentence Count:** The deviation of a tweet to sentences can depend on the meaning inside the that tweet which can also play a role in the tweet's classification.
- 4. Chars Count:** This feature calculates the number of Arabic alphabetical characters in the tweet.
- 5. Words Count:** This feature calculates the number of Arabic alphabetical words in the tweet.
- 6. Hashtags Percentage:** The number of hashtags can tell a different meaning by having different tweets lengths, so the number of hashtags to the number of the words in the tweet can help reducing this problem, so it is used here as a feature too.
- 7. Emoji Percentage:** Same as the hashtags the percentage of the included emojis to the number of words in a tweet can play in this process which can make it a feature.
- 8. Average Sentence Length:** The average length of the sentence as the summation of the number of words in each sentence to the number of sentences in a tweet is used as a feature.
- 9. Average Word Length:** The average length of the words as the summation of the Arabic characters in each word to the number of words in a tweet is used as a feature.

**10. Stop Words Count:** The number of the Arabic stop words in the tweet (it is calculated before the tweet is preprocessed since the preprocessing removes the stop words), as the number of stop words such as (يا، نحن، لولا، بعد، أي، إن، في، على، من، الذي، أكثر) can relate to the tweet's emotions.

### 2.3.2. Automated Features

The term frequency-inverse document frequency (TF-IDF) can be used in order to calculate the frequency of the emojis and Arabic words that are included in the tweets in order to extract a number of features for the tweet that can play a big role in its classification, where it can be explained in the following formulas: [4]

$$TF(t, d) = \frac{\text{number of times } t \text{ appears in } d}{\text{total number of terms in } d}$$

$$IDF(t) = \log \frac{N}{1+df}$$

$$TF - IDF(t, d) = TF(t, d) \times IDF(t)$$

### 2.4. Training & Testing

Now that the features are set, the given data set should be distributed as the system shall be trained on 75% of the given tweets (both positive and negative) by their given features, and trained on the remaining 25% by classifying them using the learnt method from the training data in order to evaluate the outcome classifications of the tested data and compare them with the original ones, figure 2-5 shows the diagram of the two processes. [5]

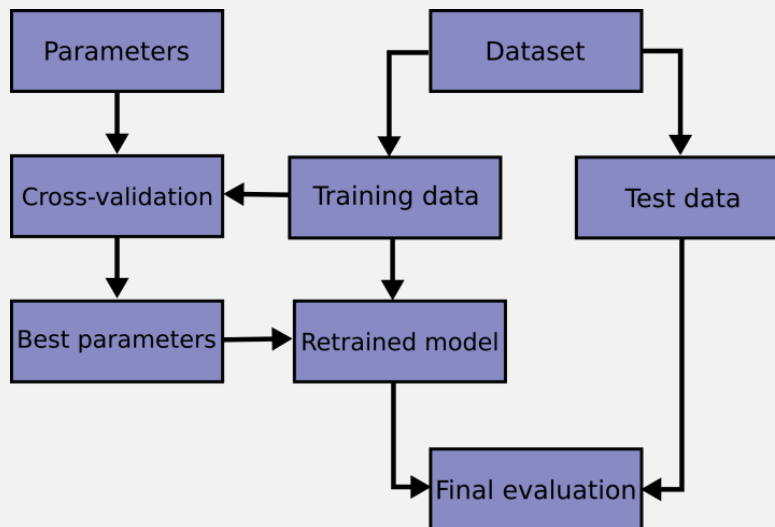


Figure 2-5: Training and Testing Diagram

## 2.5. Testing Calculations

For this project three classifiers are being used:

1. **Random Forest:** Which is an implementation of a decision tree.
2. **Naïve Bayes:** Using probability.
3. **Multi-Layer Perception:** Neural Network Classifier.

Each classifiers performance is measured by the following:

1. **Accuracy:** using the following formula

$$\frac{\text{Correct Classifications}}{\text{Total Cases}}$$

2. **Sensitivity (Recall):** using the following formula

$$\frac{\text{True Positive}}{\text{Classified Positive}}$$

3. **Precession:** using the following formula

$$\frac{\text{True Positive}}{\text{Actual Positive}}$$

4. **F1 -Score:** using the following formula

$$\frac{2 pr}{p + r}$$

5. **ROC AUC:** which is the area under the receiver operating characteristics.

In order to avoid overfitting, the average score using 5-fold cross validation is implemented, as it divides the data for training and testing 5 times, with different distribution of the testing and training data each time yet with the distribution percentage remains the same.

## 2.6. The Deployment

A set of unclassified Arabic tweets can be given to the system, after the system has learned how to classify tweets, the system will take those tweets process them and classify them by their features using the previous three classifiers, where each classifier has its own results in determining whether the tweets are positive or negative.

## 3. Running The Program

### 3.1. Data Preprocessing Results

The table below shows the out come of the preprocessing of some tweets with the Emoji extraction.

Table 3-1: Tweets Preprocessing

Tweet	Tweet's Preprocessing text	Emojis
أحببته حتى أقنعني إن ما فات العمر كان إنتظار له 🐼	احب حتي قنع ان فات عمر نظر	['🐼']
لم يبدو ان دجلة اعتادت على التهام اجساد ابنائها من سبايكر للعبارة .. 🖐️🖐️❤ ما سمعته .. ان البشر يعطش فيشرب الماء ..	بدو ان دجل عاد علي تهم جسد من يكر عبر سمع ان بشر عطش شرب ماء	['🖐️🖐️❤']
لو بيدي أرجع ساعتى وين أرجع؟ إلیا صدقه؟ والله أبقي أفر بيها لما ترجع بشر ماعرفه! وأنهاي العلاقة من العرق.. قبل الدمع	بيد رجع سعت وين رجل الا هدفه وله ابق افر بيه رجع بشر اعرف ونه علق عرق دمع	[]
انتت يمكن الي ناسي احدث البصرة قبل شهور او متغافل عنها نقتلوا متظاهرين 🙄 بدم بارد منو قتلهم	انتت يمكن الي نسي حدث بصر شهر او غافل عنه قتل ظاهر بدم بارد منو قتل	['🙄']
من يخاف فليس منا.. تصبحون على 🏠 خير #النصر الاتحاد	يخف فليس منا صبح علي خير	['🏠']

### 3.2. Feature Extraction Results

The csv that includes the processed data with the chosen features will be as in figure 3-1.

	emojiCount	hashtagsCount	sentencesCount	charsCount	wordsCount	hashtagsPercentage	emojiPercentage	avgSentenceLength	avgWordLength	stopwordsCount	tfidf_15	tfidf_16	tfidf_17	tfidf_18	tfidf_19	tfidf_20	tfidf_21	tfidf_22	tfidf_23	tfidf_24	tfidf_25	tfidf_26	tfidf_27	tfidf_28	tfidf_29	tfidf_30	tfidf_31	tfidf_32	tfidf_33	tfidf_34	tfidf_35	tfidf_36	tfidf_37	tfidf_38	tfidf_39	tfidf_40	tfidf_41	tfidf_42	tfidf_43	tfidf_44	tfidf_45	tfidf_46	tfidf_47	tfidf_48	tfidf_49	tfidf_50	tfidf_51	tfidf_52	tfidf_53	tfidf_54	tfidf_55	tfidf_56	tfidf_57	tfidf_58	tfidf_59	tfidf_60	tfidf_61	tfidf_62	tfidf_63	tfidf_64	tfidf_65	tfidf_66	tfidf_67	tfidf_68	tfidf_69	tfidf_70	tfidf_71	tfidf_72	tfidf_73	tfidf_74	tfidf_75	tfidf_76	tfidf_77	tfidf_78	tfidf_79	tfidf_80	tfidf_81	tfidf_82	tfidf_83	tfidf_84	tfidf_85	tfidf_86	tfidf_87	tfidf_88	tfidf_89	tfidf_90	tfidf_91	tfidf_92	tfidf_93	tfidf_94	tfidf_95	tfidf_96	tfidf_97	tfidf_98	tfidf_99	tfidf_100	tfidf_101	tfidf_102	tfidf_103	tfidf_104	tfidf_105	tfidf_106	tfidf_107	tfidf_108	tfidf_109	tfidf_110	tfidf_111	tfidf_112	tfidf_113	tfidf_114	tfidf_115	tfidf_116	tfidf_117	tfidf_118	tfidf_119	tfidf_120	tfidf_121	tfidf_122	tfidf_123	tfidf_124	tfidf_125	tfidf_126	tfidf_127	tfidf_128	tfidf_129	tfidf_130	tfidf_131	tfidf_132	tfidf_133	tfidf_134	tfidf_135	tfidf_136	tfidf_137	tfidf_138	tfidf_139	tfidf_140	tfidf_141	tfidf_142	tfidf_143	tfidf_144	tfidf_145	tfidf_146	tfidf_147	tfidf_148	tfidf_149	tfidf_150	tfidf_151	tfidf_152	tfidf_153	tfidf_154	tfidf_155	tfidf_156	tfidf_157	tfidf_158	tfidf_159	tfidf_160	tfidf_161	tfidf_162	tfidf_163	tfidf_164	tfidf_165	tfidf_166	tfidf_167	tfidf_168	tfidf_169	tfidf_170	tfidf_171	tfidf_172	tfidf_173	tfidf_174	tfidf_175	tfidf_176	tfidf_177	tfidf_178	tfidf_179	tfidf_180	tfidf_181	tfidf_182	tfidf_183	tfidf_184	tfidf_185	tfidf_186	tfidf_187	tfidf_188	tfidf_189	tfidf_190	tfidf_191	tfidf_192	tfidf_193	tfidf_194	tfidf_195	tfidf_196	tfidf_197	tfidf_198	tfidf_199	tfidf_200	tfidf_201	tfidf_202	tfidf_203	tfidf_204	tfidf_205	tfidf_206	tfidf_207	tfidf_208	tfidf_209	tfidf_210	tfidf_211	tfidf_212	tfidf_213	tfidf_214	tfidf_215	tfidf_216	tfidf_217	tfidf_218	tfidf_219	tfidf_220	tfidf_221	tfidf_222	tfidf_223	tfidf_224	tfidf_225	tfidf_226	tfidf_227	tfidf_228	tfidf_229	tfidf_230	tfidf_231	tfidf_232	tfidf_233	tfidf_234	tfidf_235	tfidf_236	tfidf_237	tfidf_238	tfidf_239	tfidf_240	tfidf_241	tfidf_242	tfidf_243	tfidf_244	tfidf_245	tfidf_246	tfidf_247	tfidf_248	tfidf_249	tfidf_250	tfidf_251	tfidf_252	tfidf_253	tfidf_254	tfidf_255	tfidf_256	tfidf_257	tfidf_258	tfidf_259	tfidf_260	tfidf_261	tfidf_262	tfidf_263	tfidf_264	tfidf_265	tfidf_266	tfidf_267	tfidf_268	tfidf_269	tfidf_270	tfidf_271	tfidf_272	tfidf_273	tfidf_274	tfidf_275	tfidf_276	tfidf_277	tfidf_278	tfidf_279	tfidf_280	tfidf_281	tfidf_282	tfidf_283	tfidf_284	tfidf_285	tfidf_286	tfidf_287	tfidf_288	tfidf_289	tfidf_290	tfidf_291	tfidf_292	tfidf_293	tfidf_294	tfidf_295	tfidf_296	tfidf_297	tfidf_298	tfidf_299	tfidf_300	tfidf_301	tfidf_302	tfidf_303	tfidf_304	tfidf_305	tfidf_306	tfidf_307	tfidf_308	tfidf_309	tfidf_310	tfidf_311	tfidf_312	tfidf_313	tfidf_314	tfidf_315	tfidf_316	tfidf_317	tfidf_318	tfidf_319	tfidf_320	tfidf_321	tfidf_322	tfidf_323	tfidf_324	tfidf_325	tfidf_326	tfidf_327	tfidf_328	tfidf_329	tfidf_330	tfidf_331	tfidf_332	tfidf_333	tfidf_334	tfidf_335	tfidf_336	tfidf_337	tfidf_338	tfidf_339	tfidf_340	tfidf_341	tfidf_342	tfidf_343	tfidf_344	tfidf_345	tfidf_346	tfidf_347	tfidf_348	tfidf_349	tfidf_350	tfidf_351	tfidf_352	tfidf_353	tfidf_354	tfidf_355	tfidf_356	tfidf_357	tfidf_358	tfidf_359	tfidf_360	tfidf_361	tfidf_362	tfidf_363	tfidf_364	tfidf_365	tfidf_366	tfidf_367	tfidf_368	tfidf_369	tfidf_370	tfidf_371	tfidf_372	tfidf_373	tfidf_374	tfidf_375	tfidf_376	tfidf_377	tfidf_378	tfidf_379	tfidf_380	tfidf_381	tfidf_382	tfidf_383	tfidf_384	tfidf_385	tfidf_386	tfidf_387	tfidf_388	tfidf_389	tfidf_390	tfidf_391	tfidf_392	tfidf_393	tfidf_394	tfidf_395	tfidf_396	tfidf_397	tfidf_398	tfidf_399	tfidf_400	tfidf_401	tfidf_402	tfidf_403	tfidf_404	tfidf_405	tfidf_406	tfidf_407	tfidf_408	tfidf_409	tfidf_410	tfidf_411	tfidf_412	tfidf_413	tfidf_414	tfidf_415	tfidf_416	tfidf_417	tfidf_418	tfidf_419	tfidf_420	tfidf_421	tfidf_422	tfidf_423	tfidf_424	tfidf_425	tfidf_426	tfidf_427	tfidf_428	tfidf_429	tfidf_430	tfidf_431	tfidf_432	tfidf_433	tfidf_434	tfidf_435	tfidf_436	tfidf_437	tfidf_438	tfidf_439	tfidf_440	tfidf_441	tfidf_442	tfidf_443	tfidf_444	tfidf_445	tfidf_446	tfidf_447	tfidf_448	tfidf_449	tfidf_450	tfidf_451	tfidf_452	tfidf_453	tfidf_454	tfidf_455	tfidf_456	tfidf_457	tfidf_458	tfidf_459	tfidf_460	tfidf_461	tfidf_462	tfidf_463	tfidf_464	tfidf_465	tfidf_466	tfidf_467	tfidf_468	tfidf_469	tfidf_470	tfidf_471	tfidf_472	tfidf_473	tfidf_474	tfidf_475	tfidf_476	tfidf_477	tfidf_478	tfidf_479	tfidf_480	tfidf_481	tfidf_482	tfidf_483	tfidf_484	tfidf_485	tfidf_486	tfidf_487	tfidf_488	tfidf_489	tfidf_490	tfidf_491	tfidf_492	tfidf_493	tfidf_494	tfidf_495	tfidf_496	tfidf_497	tfidf_498	tfidf_499	tfidf_500	tfidf_501	tfidf_502	tfidf_503	tfidf_504	tfidf_505	tfidf_506	tfidf_507	tfidf_508	tfidf_509	tfidf_510	tfidf_511	tfidf_512	tfidf_513	tfidf_514	tfidf_515	tfidf_516	tfidf_517	tfidf_518	tfidf_519	tfidf_520	tfidf_521	tfidf_522	tfidf_523	tfidf_524	tfidf_525	tfidf_526	tfidf_527	tfidf_528	tfidf_529	tfidf_530	tfidf_531	tfidf_532	tfidf_533	tfidf_534	tfidf_535	tfidf_536	tfidf_537	tfidf_538	tfidf_539	tfidf_540	tfidf_541	tfidf_542	tfidf_543	tfidf_544	tfidf_545	tfidf_546	tfidf_547	tfidf_548	tfidf_549	tfidf_550	tfidf_551	tfidf_552	tfidf_553	tfidf_554	tfidf_555	tfidf_556	tfidf_557	tfidf_558	tfidf_559	tfidf_560	tfidf_561	tfidf_562	tfidf_563	tfidf_564	tfidf_565	tfidf_566	tfidf_567	tfidf_568	tfidf_569	tfidf_570	tfidf_571	tfidf_572	tfidf_573	tfidf_574	tfidf_575	tfidf_576	tfidf_577	tfidf_578	tfidf_579	tfidf_580	tfidf_581	tfidf_582	tfidf_583	tfidf_584	tfidf_585	tfidf_586	tfidf_587	tfidf_588	tfidf_589	tfidf_590	tfidf_591	tfidf_592	tfidf_593	tfidf_594	tfidf_595	tfidf_596	tfidf_597	tfidf_598	tfidf_599	tfidf_600	tfidf_601	tfidf_602	tfidf_603	tfidf_604	tfidf_605	tfidf_606	tfidf_607	tfidf_608	tfidf_609	tfidf_610	tfidf_611	tfidf_612	tfidf_613	tfidf_614	tfidf_615	tfidf_616	tfidf_617	tfidf_618	tfidf_619	tfidf_620	tfidf_621	tfidf_622	tfidf_623	tfidf_624	tfidf_625	tfidf_626	tfidf_627	tfidf_628	tfidf_629	tfidf_630	tfidf_631	tfidf_632	tfidf_633	tfidf_634	tfidf_635	tfidf_636	tfidf_637	tfidf_638	tfidf_639	tfidf_640	tfidf_641	tfidf_642	tfidf_643	tfidf_644	tfidf_645	tfidf_646	tfidf_647	tfidf_648	tfidf_649	tfidf_650	tfidf_651	tfidf_652	tfidf_653	tfidf_654	tfidf_655	tfidf_656	tfidf_657	tfidf_658	tfidf_659	tfidf_660	tfidf_661	tfidf_662	tfidf_663	tfidf_664	tfidf_665	tfidf_666	tfidf_667	tfidf_668	tfidf_669	tfidf_670	tfidf_671	tfidf_672	tfidf_673	tfidf_674	tfidf_675	tfidf_676	tfidf_677	tfidf_678	tfidf_679	tfidf_680	tfidf_681	tfidf_682	tfidf_683	tfidf_684	tfidf_685	tfidf_686	tfidf_687	tfidf_688	tfidf_689	tfidf_690	tfidf_691	tfidf_692	tfidf_693	tfidf_694	tfidf_695	tfidf_696	tfidf_697	tfidf_698	tfidf_699	tfidf_700	tfidf_701	tfidf_702	tfidf_703	tfidf_704	tfidf_705	tfidf_706	tfidf_707	tfidf_708	tfidf_709	tfidf_710	tfidf_711	tfidf_712	tfidf_713	tfidf_714	tfidf_715	tfidf_716	tfidf_717	tfidf_718	tfidf_719	tfidf_720	tfidf_721	tfidf_722	tfidf_723	tfidf_724	tfidf_725	tfidf_726	tfidf_727	tfidf_728	tfidf_729	tfidf_730	tfidf_731	tfidf_732	tfidf_733	tfidf_734	tfidf_735	tfidf_736	tfidf_737	tfidf_738	tfidf_739	tfidf_740	tfidf_741	tfidf_742	tfidf_743	tfidf_744	tfidf_745	tfidf_746	tfidf_747	tfidf_748	tfidf_749	tfidf_750	tfidf_751	tfidf_752	tfidf_753	tfidf_754	tfidf_755	tfidf_756	tfidf_757	tfidf_758	tfidf_759	tfidf_760	tfidf_761	tfidf_762	tfidf_763	tfidf_764	tfidf_765	tfidf_766	tfidf_767	tfidf_768	tfidf_769	tfidf_770	tfidf_771	tfidf_772	tfidf_773	tfidf_774	tfidf_775	tfidf_776	tfidf_777	tfidf_778	tfidf_779	tfidf_780	tfidf_781	tfidf_782	tfidf_783	tfidf_784	tfidf_785	tfidf_786	tfidf_787	tfidf_788	tfidf_789	tfidf_790	tfidf_791	tfidf_792	tfidf_793	tfidf_794	tfidf_795	tfidf_796	tfidf_797	tfidf_798	tfidf_799	tfidf_800	tfidf_801	tfidf_802	tfidf_803	tfidf_804	tfidf_805	tfidf_806	tfidf_807	tfidf_808	tfidf_809	tfidf_810	tfidf_811	tfidf_812	tfidf_813	tfidf_814	tfidf_815	tfidf_816	tfidf_817	tfidf_818	tfidf_819	tfidf_820	tfidf_821	tfidf_822	tfidf_823	tfidf_824	tfidf_825	tfidf_826	tfidf_827	tfidf_828	tfidf_829	tfidf_830	tfidf_831	tfidf_832	tfidf_833	tfidf_834	tfidf_835	tfidf_836	tfidf_837	tfidf_838	tfidf_839	tfidf_840	tfidf_841	tfidf_842	tfidf_843	tfidf_844	tfidf_845	tfidf_846	tfidf_847	tfidf_848	tfidf_849	tfidf_850	tfidf_851	tfidf_852	tfidf_853	tfidf_854	tfidf_855	tfidf_856	tfidf_857	tfidf_858	tfidf_859	tfidf_860	tfidf_861	tfidf_862	tfidf_863	tfidf_864	tfidf_865	tfidf_866	tfidf_867	tfidf_868	tfidf_869	tfidf_870	tfidf_871	tfidf_872	tfidf_873	tfidf_874	tfidf_875	tfidf_876	tfidf_877	tfidf_878	tfidf_879	tfidf_880	tfidf_881	tfidf_882	tfidf_883	tfidf_884	tfidf_885	tfidf_886	tfidf_887	tfidf_888	tfidf_889	tfidf_890	tfidf_891	tfidf_892	tfidf_893	tfidf_894	tfidf_895	tfidf_896	tfidf_897	tfidf_898	tfidf_899	tfidf_900	tfidf_901	tfidf_902	tfidf_903	tfidf_904	tfidf_905	tfidf_906	tfidf_907	tfidf_908	tfidf_909	tfidf_910	tfidf_911	tfidf_912	tfidf_913	tfidf_914	tfidf_915	tfidf_916	tfidf_917	tfidf_918	tfidf_919	tfidf_920	tfidf_921	tfidf_922	tfidf_923	tfidf_924	tfidf_925	tfidf_926	tfidf_927	tfidf_928	tfidf_929	tfidf_930	tfidf_931	tfidf_932	tfidf_933	tfidf_934	tfidf_935	tfidf_936	tfidf_937	tfidf_938	tfidf_939	tfidf_940	tfidf_941	tfidf_942	tfidf_943	tfidf_944	tfidf_945	tfidf_946	tfidf_947	tfidf_948	tfidf_949	tfidf_950	tfidf_951	tfidf_952	tfidf_953	tfidf_954	tfidf_955	tfidf_956	tfidf_957	tfidf_958	tfidf_959	tfidf_960	tfidf_961	tfidf_962	tfidf_963	tfidf_964	tfidf_965	tfidf_966	tfidf_967	tfidf_968	tfidf_969	tfidf_970	tfidf_971	tfidf_972	tfidf_973	tfidf_974	tfidf_975	tfidf_976	tfidf_977	tfidf_978	tfidf_979	tfidf_980	tfidf_981	tfidf_982	tfidf_983	tfidf_984	tfidf_985	tfidf_986	tfidf_987	tfidf_988	tfidf_989	tfidf_990	tfidf_991	tfidf_992	tfidf_993	tfidf_994	tfidf_995	tfidf_996	tfidf_997	tfidf_998	tfidf_999	tfidf_1000	tfidf_1001	tfidf_1002	tfidf_1003	tfidf_1004	tfidf_1005	tfidf_1006	tfidf_1007	tfidf_1008	tfidf_1009	tfidf_1010	tfidf_1011	tfidf_1012	tfidf_1013	tfidf_1014	tfidf_1015	tfidf_1016	tfidf_1017	tfidf_1018	tfidf_1019	tfidf_1020	tfidf_1021	tfidf_1022	tfidf_1023	tfidf_1024	tfidf_1025	tfidf_1026	tfidf_1027	tfidf_1028	tfidf_1029	tfidf_1030	tfidf_1031	tfidf_1032	tfidf_1033	tfidf_1034	tfidf_1035	tfidf_1036	tfidf_1037	tfidf_1038	tfidf_1039	tfidf_1040	tfidf_1041	tfidf_1042	tfidf_1043	tfidf_1044	tfidf_1045	tfidf_1046	tfidf_1047	tfidf_1048	tfidf_1049	tfidf_1050	tfidf_1051	tfidf_1052	tfidf_1053	tfidf_1054	tfidf_1055	tfidf_1056	tfidf_1057	tfidf_1058	tfidf_1059	tfidf_1060	tfidf_1061	tfidf_1062	tfidf_1063
--	------------	---------------	----------------	------------	------------	--------------------	-----------------	-------------------	---------------	----------------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	-----------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------	------------

Each features effect is calculated and plotted as a histogram where the it gives the average value of each feature for both classifications, as its all shown below.

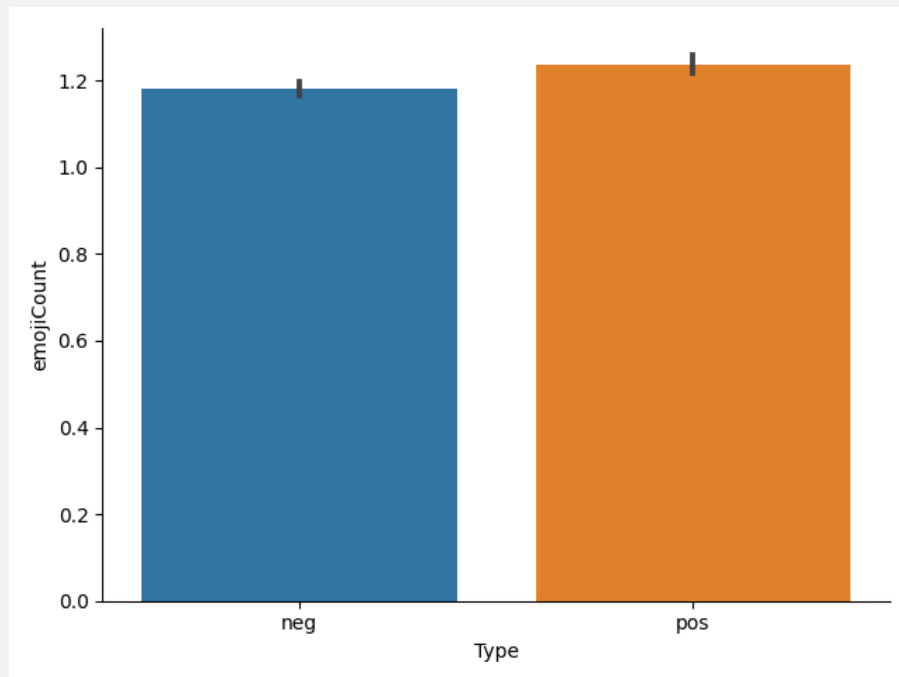


Figure 3-2: Emoji Counting Difference

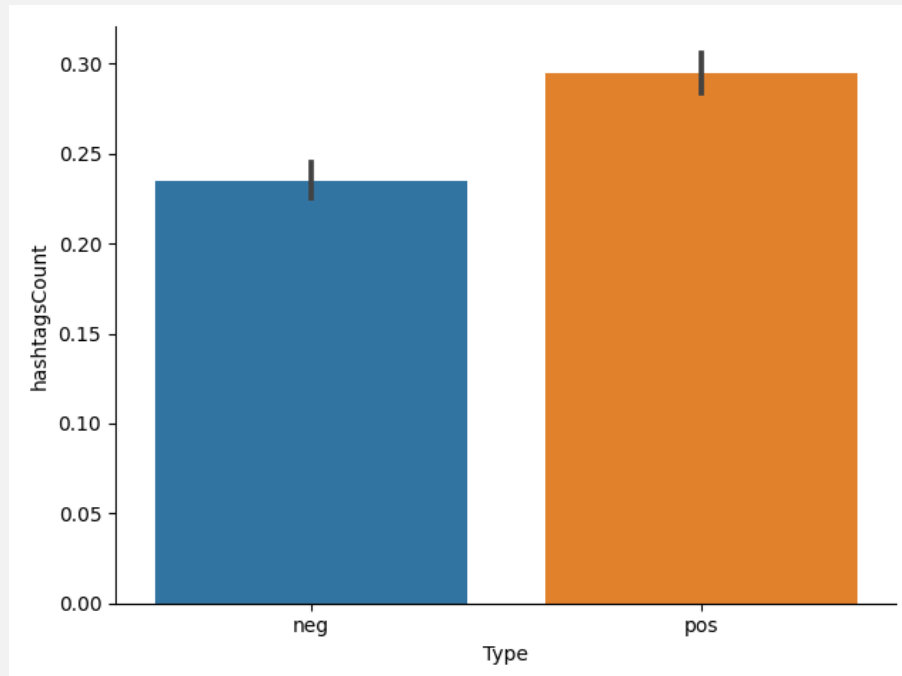
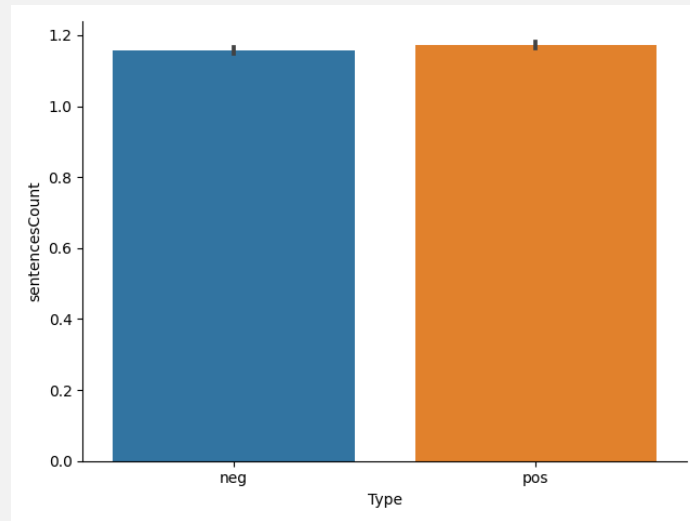
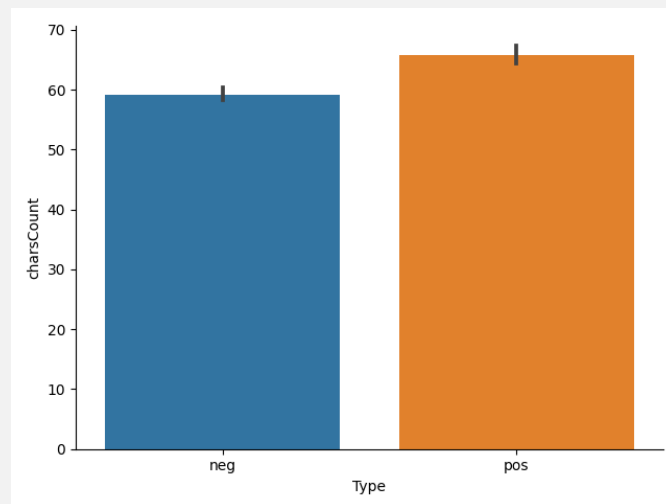


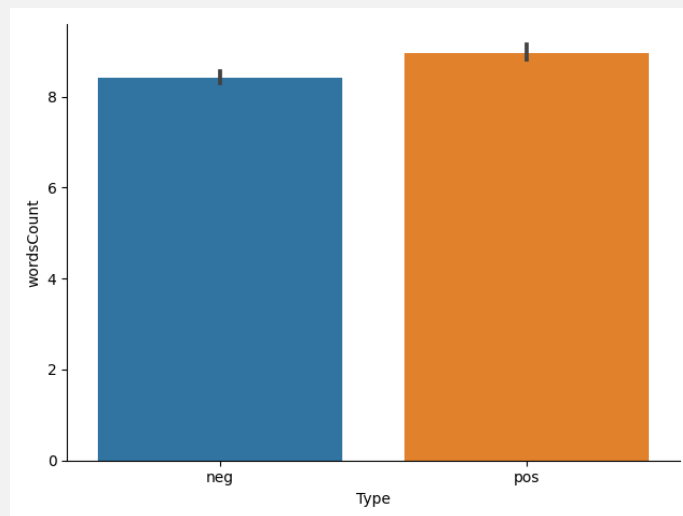
Figure 3-3: Hashtag Counting Difference



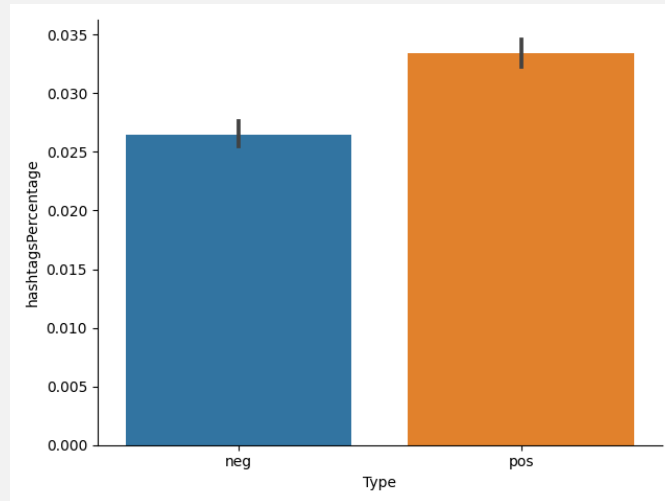
**Figure 3-4: Sentence Counting Difference**



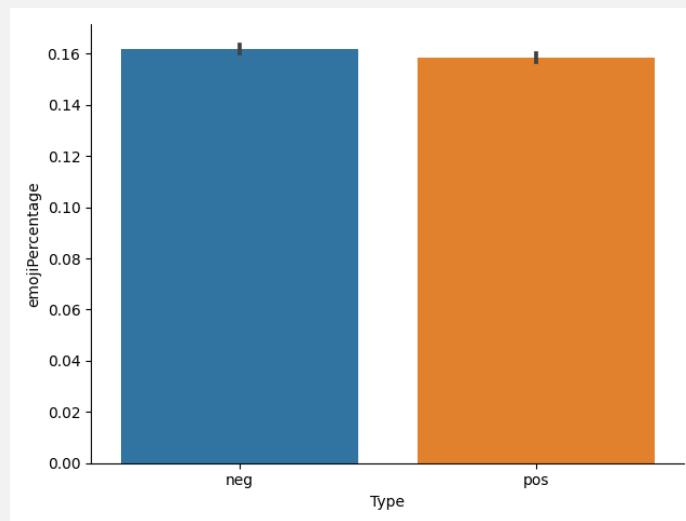
**Figure 3-5: Character Counting Difference**



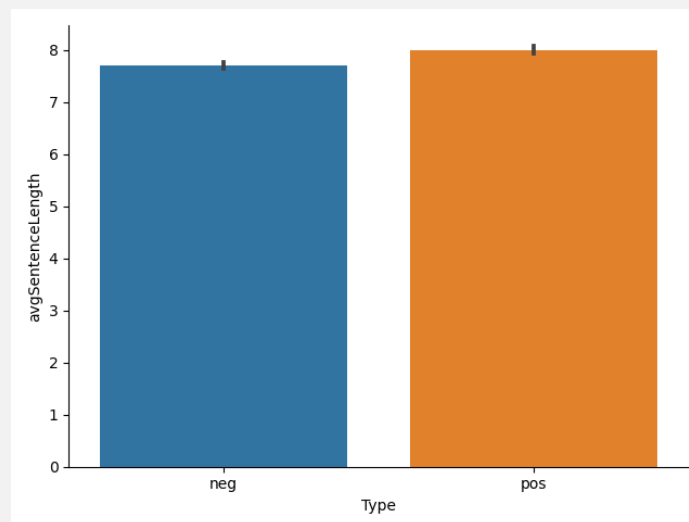
**Figure 3-6: Word Counting Difference**



**Figure 3-7: Hashtag Percentage Difference**



**Figure 3-8: Emoji Percentage Difference**



**Figure 3-9: Average words in a sentence Difference**

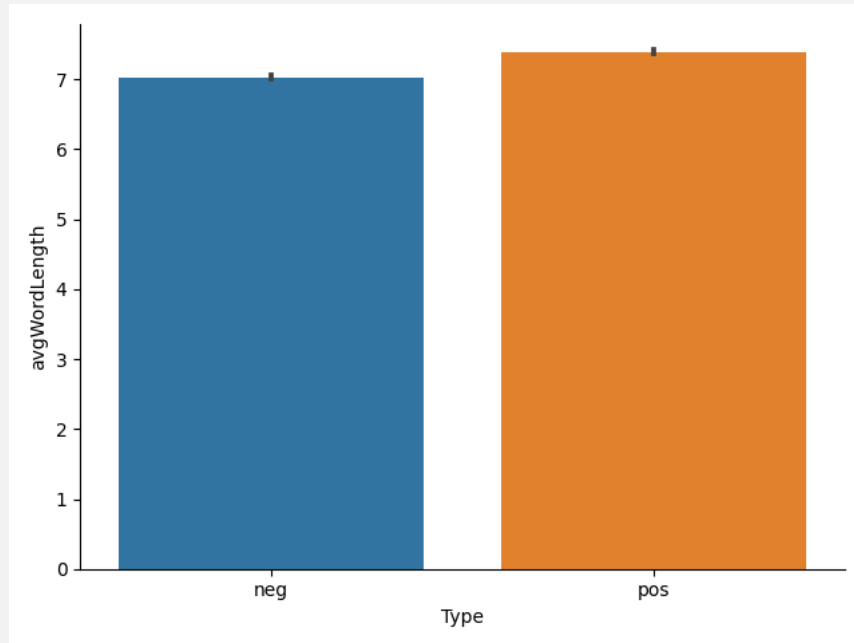


Figure 3-10: Average Characters in a Word Difference

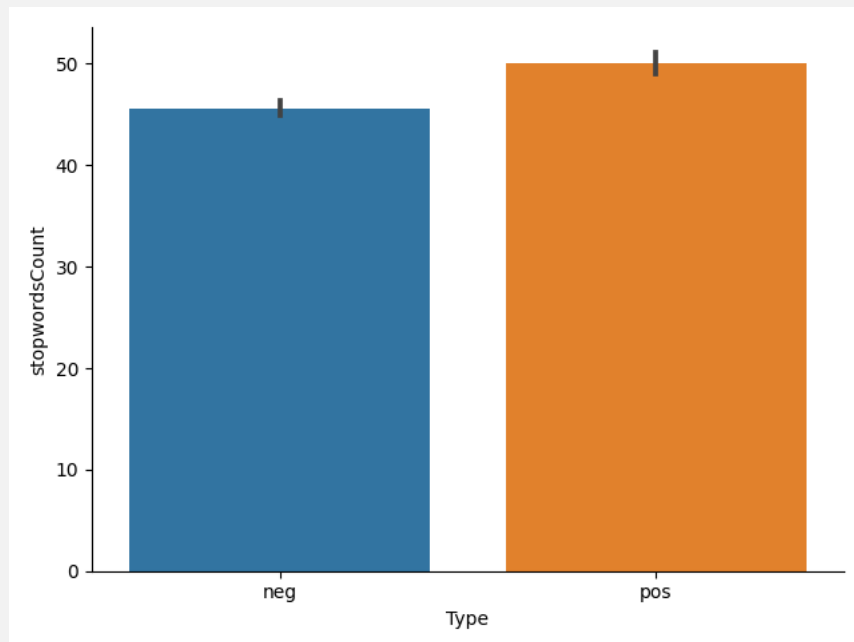


Figure 3-11: Stop Words Counting Difference



### 3.3. Classifications Models Results

Each classifier model has its confusion matrix, and precession, recall, accuracy, and F1-rate, adding on the ROC AUC and the scores using 5-fold cross validation.

#### 3.3.1. Random Forest

The Confusion matrix for the random forest classifications is displayed in figure 3-12, with the true positive rate graph in figure 3-13, false negative rate graph in figure 3-14, and the precision recall graph in figure 3-15.

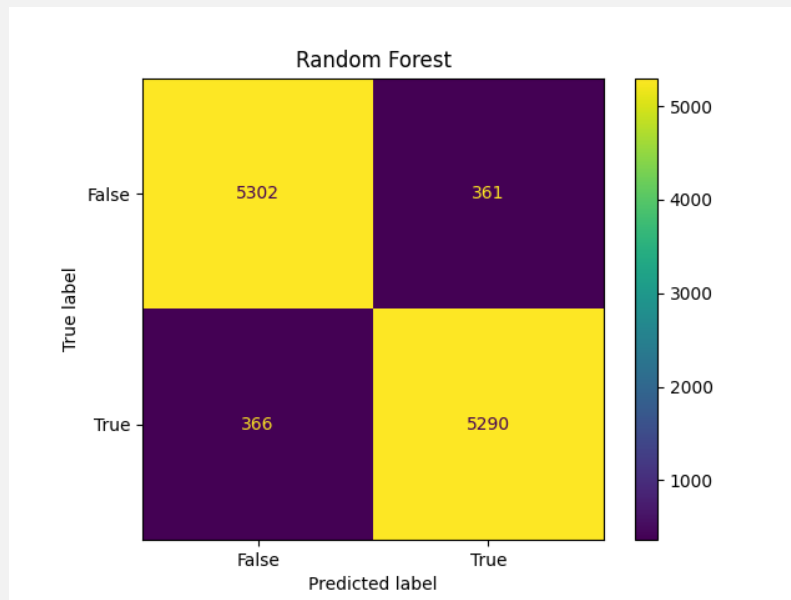


Figure 3-12: Random Forest Confusion Matrix

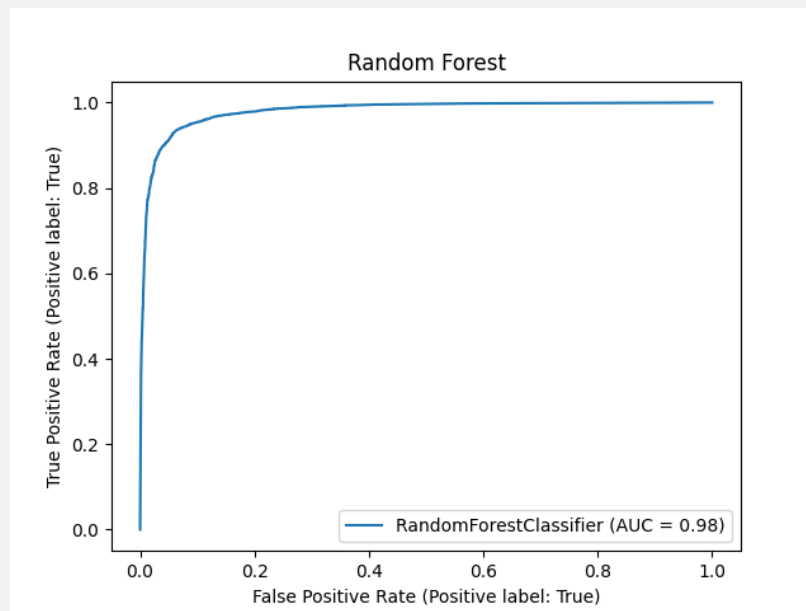
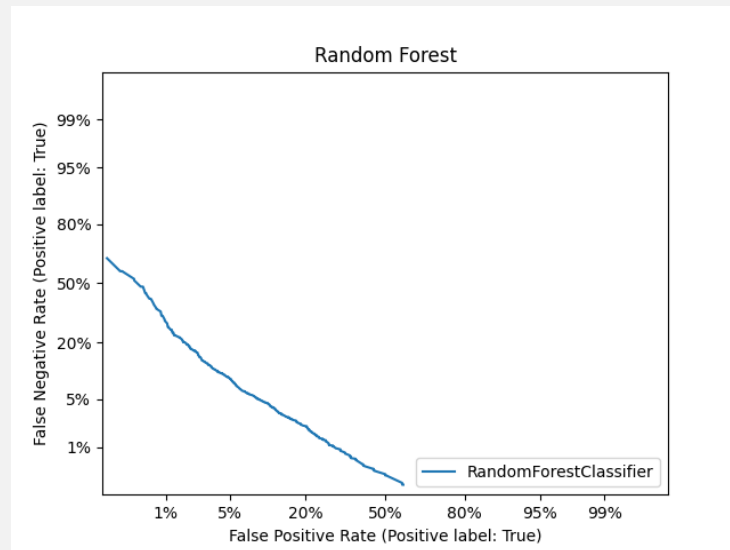
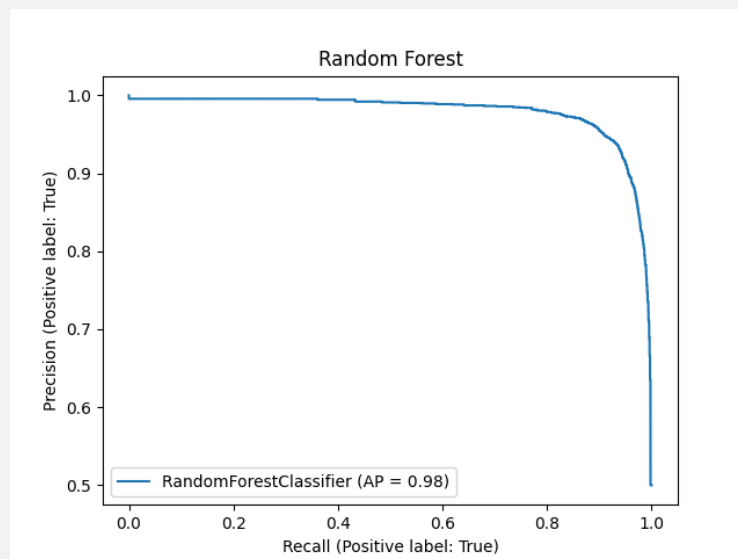


Figure 3-13: Random Forest True Positive Graph



**Figure 3-14: Random Forest False Negative Graph**



**Figure 3-15: Random Forest Precision-Recall Graph**

The Performance measurements for the random forest were as follows:

**Sensitivity (recall) score:** 0.9352899575671852

**precision score:** 0.9361175013271987

**f1 score:** 0.9357035464756346

**accuracy score:** 0.9357717112819154

**ROC AUC:** 0.9799541385169519

**Scores using 5-fold cross validation:** [0.9326339 0.93175041 0.93705135 0.93219216 0.92932082]

**Average score using 5-fold cross validation:** 0.933

### 3.3.2. Bernoulli Naïve Bayes

The Confusion matrix for the Naïve Bayes classifications is displayed in figure 3-16, with the true positive rate graph in figure 3-17, false negative rate graph in figure 3-18, and the precision recall graph in figure 3-19.

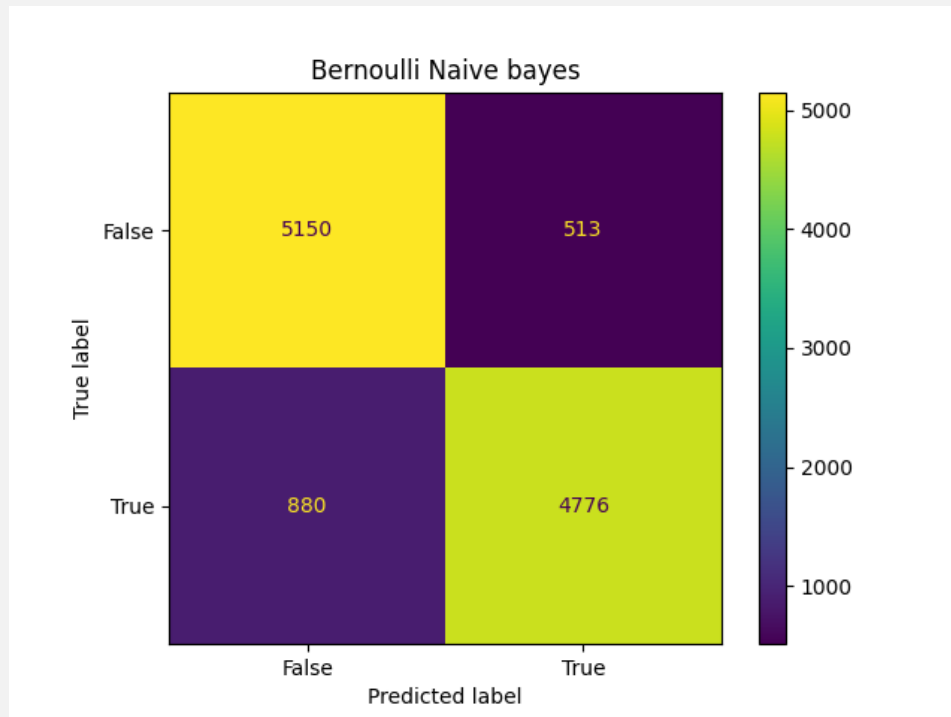


Figure 3-16: Naïve Bayes Confusion Matrix

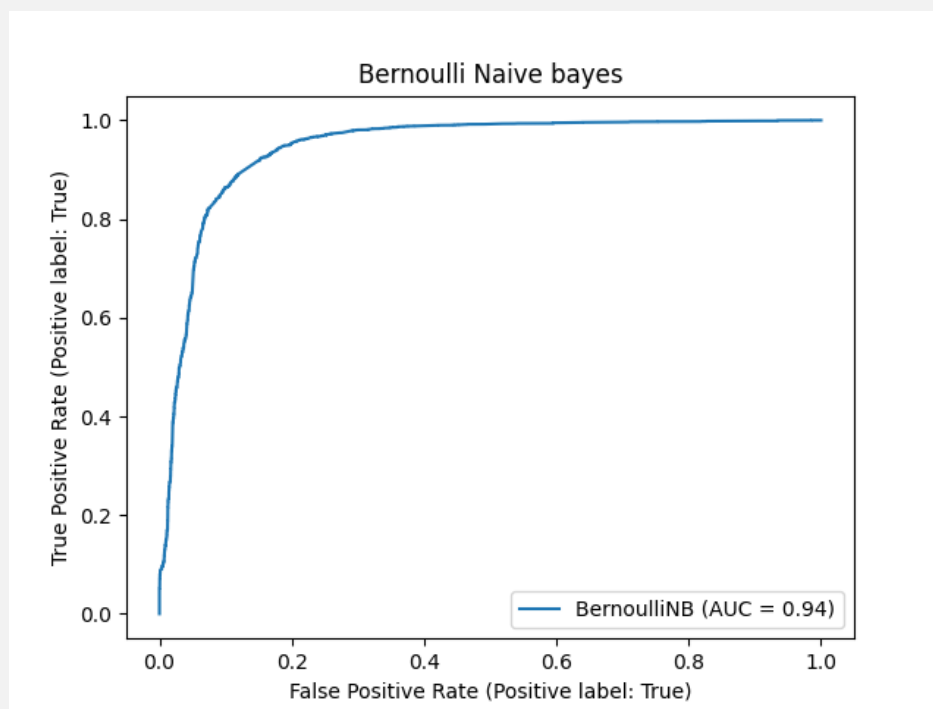


Figure 3-17: Naïve Bayes True Positive Graph

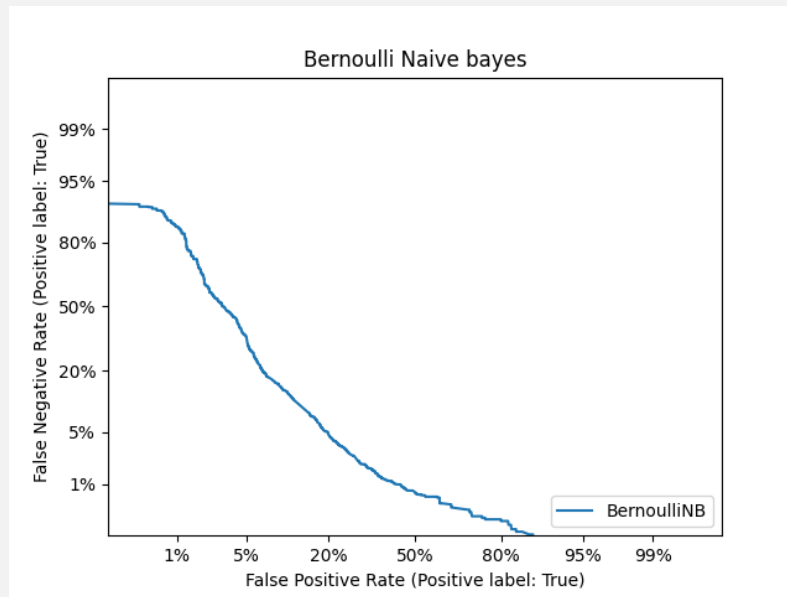


Figure 3-18: Naïve Bayes False Negative Graph

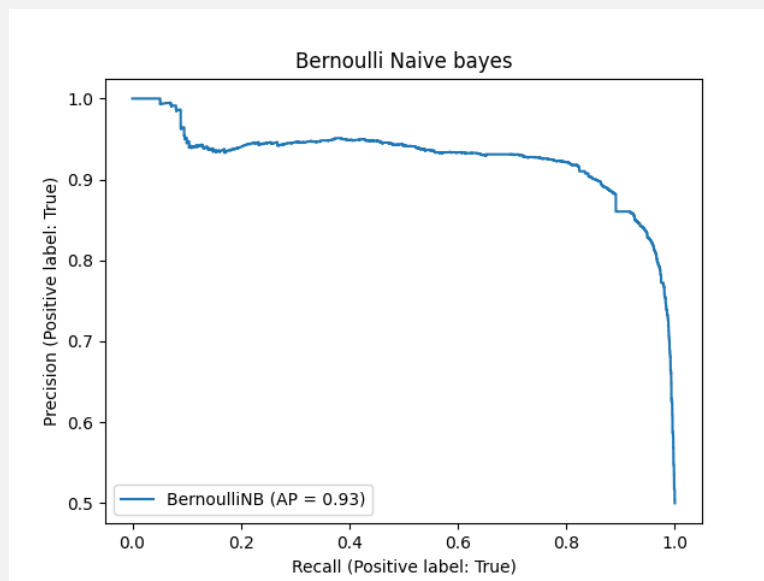


Figure 3-19: Naïve Bayes Precision-Recall Graph

The Performance measurements for naïve Bayes were as follows:

**Sensitivity (recall) score:** 0.8444130127298444

**precision score:** 0.9030062393647192

**f1 score:** 0.8727272727272726

**accuracy score:** 0.8769325912183055

**ROC AUC:** 0.9440795183804347

**Scores using 5-fold cross validation:** [0.87818885 0.8725566 0.8753175 0.87410271 0.87553838]

**Average score using 5-fold cross validation:** 0.875

### 3.3.3. Multi-Layer Perceptron

The Confusion matrix for the Multi-Layer Perceptron classifications is displayed in figure 3-20, with the true positive rate graph in figure 3-21, false negative rate graph in figure 3-22.

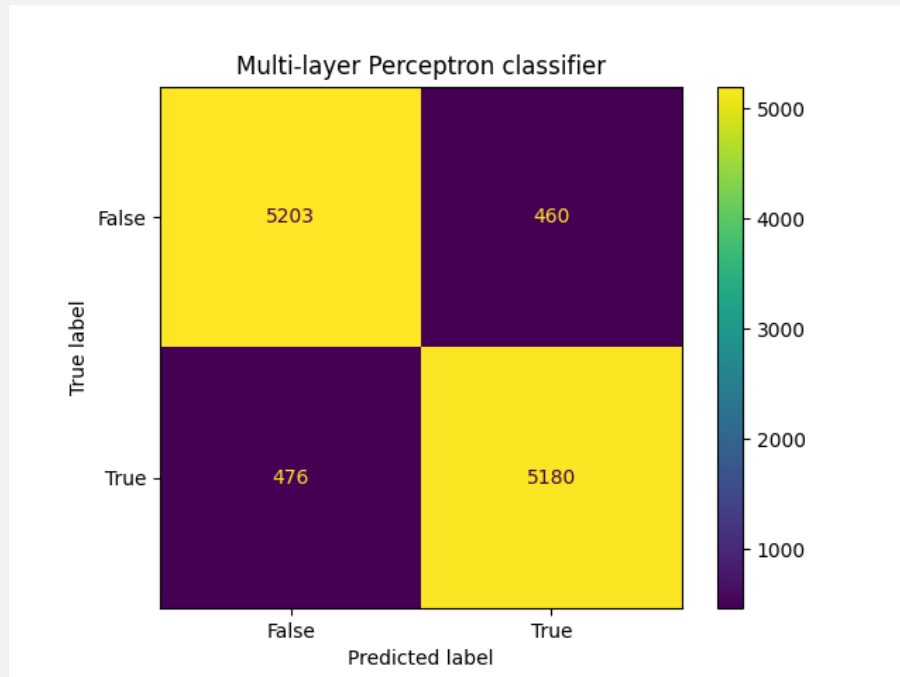


Figure 3-20: Multi-Layer Perceptron Confusion Matrix

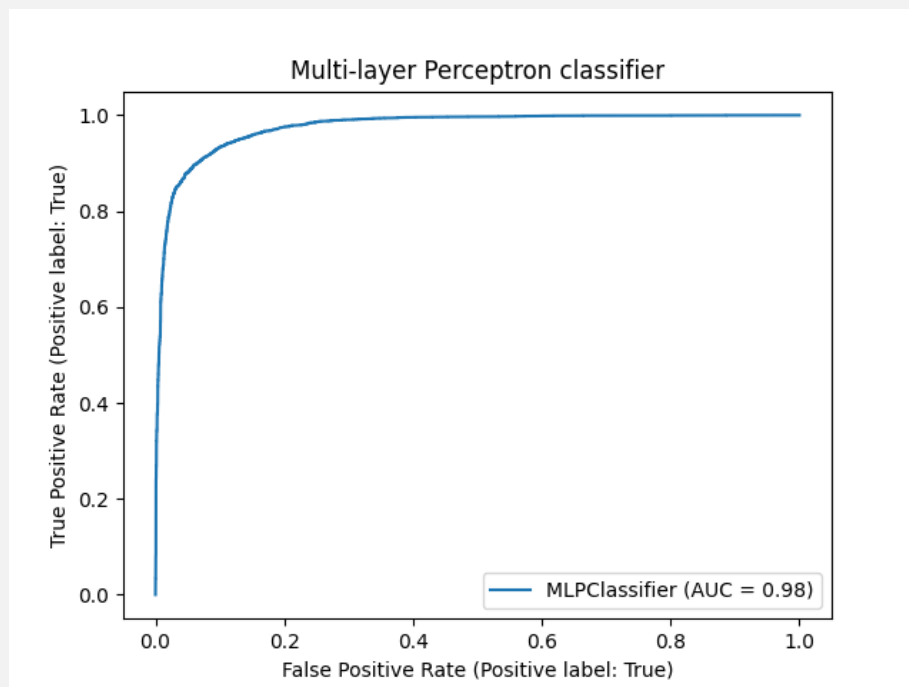


Figure 3-21: Multi-Layer Perceptron True Positive Graph

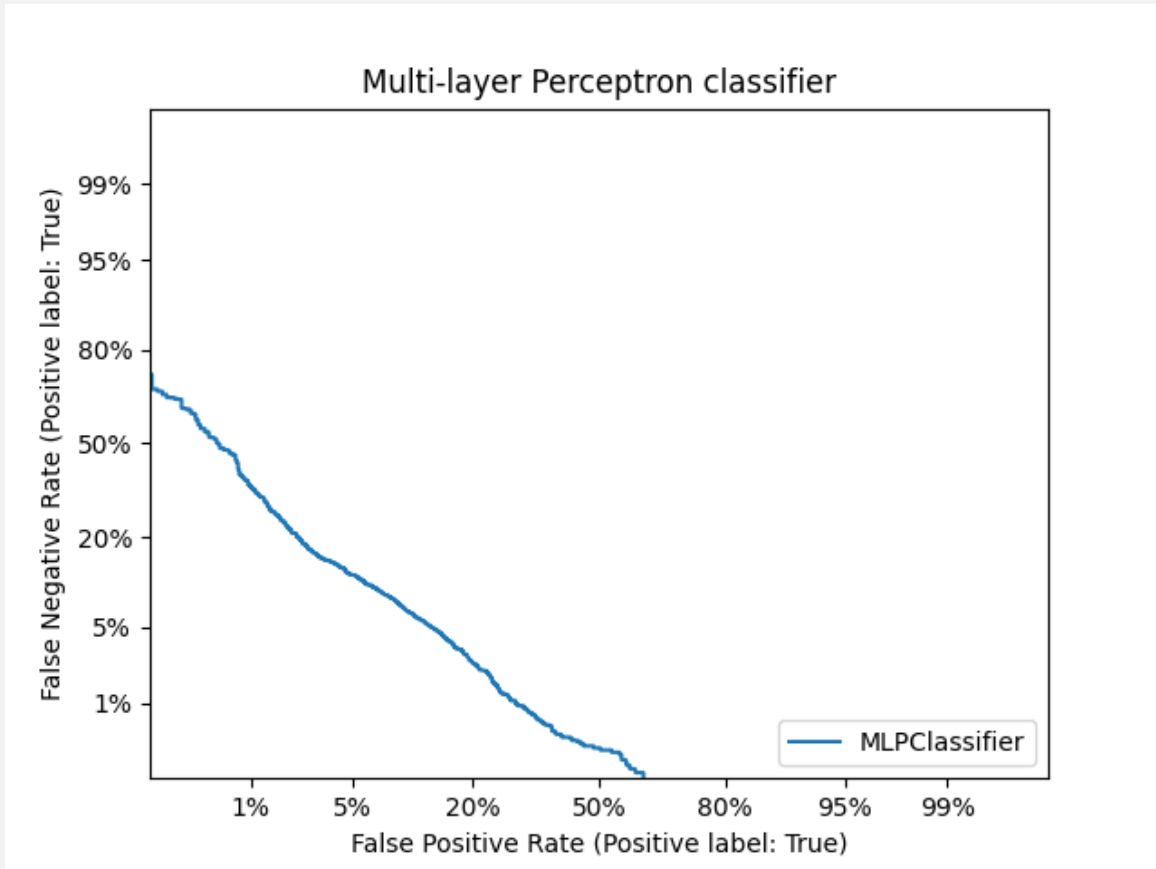


Figure 3-21: Multi-Layer Perceptron False Negative Graph

The Performance measurements for the Multi-Layer Perceptron were as follows:

**Sensitivity (recall) score:** 0.9158415841584159

**precision score:** 0.9184397163120568

**f1 score:** 0.9171388101983002

**accuracy score:** 0.917307182613305

**ROC AUC:** 0.9751695976338131

**Scores using 5-fold cross validation:** [0.92269464 0.92203203 0.91728327 0.91109884 0.91540585]

**Average score using 5-fold cross validation:** 0.918

### 3.3.4. Comparing Between Classifiers

First it is noticed that the accuracy's highest value was less than 94%, and that's because most of the features are extracted from the emojis and few emojis were extracted from the text. In order to get better performance of the model, useful features must be extracted from the text.

The performance difference in the three classifications can be found in table 3-2 below.

**Table 3-2: Classifications Performance Comparison**

Classification	Accuracy	Precision	Recall	F1	ROC AUC	Avg using 5-fold cross validation
Random Forest	0.936	0.936	0.935	0.936	0.98	0.933
Naïve Bayes	0.878	0.903	0.844	0.873	0.944	0.875
Multi-Layer Perceptron	0.917	0.918	0.916	0.917	0.975	0.918

As seen in table 3-2 the performance measurements for the random forest were very close, same thing for the Multi-layer perceptron, yet the Naïve Bayes classification gave 90% precision rate, yet 84% recall rate.

From the table it is obvious that the random forest has resulted in the best performance of all the three classifications as it gave the best precision, recall, and ROC AUC, where the Multi-layer Perceptron came in second, and Naïve Bayes came in last.

## 4. Conclusion

In conclusion after working on this project, a better understanding of machine learning mechanism is achieved, as it showed how computers take the smallest details in the data set and converts them in short time into classifications and decisions.

this project also gives a better understanding of the importance and the needs of its stages, where data preprocessing was implemented in order to give the computer a clear data that can result in giving more accurate classifications, feature extraction shows what are the important and needed things in the data for its classification, training is what teaches the system to become familiar with the data, and testing tells the situation of the entire systems quality.

Machine learning and natural language processing are currently playing a big role in the world today as most likely everything depends on them in either (banks, governments, medicine, social medic, etc..), as they made life easier and saved peoples time in data gathering and making decisions, as the computer does things in a way faster time.



## 5. References

[1]. Arabic words vocalization – arabDict.com:

<https://www.arabdict.com/en/english-arabic>

Accessed on February 15th 2023, at 12:01 AM

[2]. Stemming – researchgate.net:

[https://www.researchgate.net/publication/320607773\\_Arabic\\_information\\_retrieval\\_Stemming\\_or\\_lemmatization](https://www.researchgate.net/publication/320607773_Arabic_information_retrieval_Stemming_or_lemmatization)

Accessed on February 15th 2023, at 12:40 AM

[3]. Feature Extraction – medium.com:

<https://medium.com/analytics-vidhya/feature-extraction-and-challenges-a1e4f3f4cb53>

Accessed on February 17th 2023, at 1:40 AM

[4]. TF-IDF –kinder-chen.medium.com:

<https://kinder-chen.medium.com/introduction-to-natural-language-processing-tf-idf-1507e907c19>

Accessed on February 17th 2023, at 2:27 AM

[5]. Training and Testing – researchgate.net:

[https://www.researchgate.net/figure/Training-and-testing-our-machine-learning-approach\\_fig2\\_318132501](https://www.researchgate.net/figure/Training-and-testing-our-machine-learning-approach_fig2_318132501)

Accessed on February 17th 2023, at 3:14 AM