ICS-490-01: Special Topics I (T171)

Dr. Lahouari Ghouti

Twitter Dataset Report

*Group Members*

|  |  |
| --- | --- |
| **Ibrahim BinAlshikh** | 201139750 |
| **Ibrahim Beladi** | 201224780 |
| **Mohammed Alfarraj** | 201229820 |
| **Omar Al-Dakhil** | 201381890 |

**Abstract**

Recently, Saudi Government allowed Cinema in KSA and there were many differing opinions on twitter. Cinema in KSA or #السينما\_في\_السعودية contains many tweets between supporters and opposers. Tweets are collected by python code and stored on csv file. Next, we had to label them by four labels: “Positive” , “Negative” , “ neutral “ , “ spam”. Then had to get some important words and then finally used them on classifiers.

**Table of Contents**

[**Toolbox**](#_ifusqwomc5g8) **2**

[**Machine Learning Solution**](#_vq3sa87a2wbb) **2**

[Algorithm Description](#_2v1tqx5ynilb) 2

[Performance Measures Used](#_i6o30oht21tt) 2

[**Exploratory Data Analysis**](#_s3gapp60i22s) **2**

[**Procedure**](#_3nco3mu6eu5p) **3**

[Data Gathering](#_5h2ozzm5jq4w) 3

[Data Labeling](#_r71jpp6hnmhf) 3

[Data Cleaning](#_sm9r28s6m8no) 4

[Features Extraction](#_3lnkq2coxha4) 4

[Classifiers](#_mdldmyj3nz8c) 4

[**Performance Results**](#_aixgyqjo5bsc) **5**

[Support Vector Machines Classifier](#_s7q4qwh1cb03) 5

[Multinomial Naïve Bayes Classifier](#_wly73c5zgq17) 6

[Random Forests Classifier](#_n4i1to8l5l3j) 7

[Logistic Regression Classifier](#_dla3az55m8fs) 8

[Voting Classifier](#_xuuj2vphfn83) 9

[**Parameters**](#_ib5ang590t1j) **10**

[Support Vector Machines Classifier](#_2nemgn7r3zlm) 10

[Multinomial Naïve Bayes Classifier](#_xasqyiq8n8mt) 10

[Random Forests Classifier](#_thw4o1uxol3n) 10

[Logistic Regression Classifier](#_wo2yhhicubfq) 10

[Voting Classifier](#_k5ou8v3jzjgp) 10

[**Tweets’ Geological Distribution**](#_gqr8idtkxy9x) **11**

[**Conclusion**](#_tjxd2r31fmvt) **11**

# 

# 

# 

# Toolbox

* Twitter APIs
* Scikit
* NumPy
* Pandas
* Matplotlib
* Folium
* Jupyter

# 

# Machine Learning Solution

## Algorithm Description

* Support vector machines.
* Multinomial Naïve Bayes.
* Random forests.
* Logistic Regression Classifier

## Performance Measures Used

* Accuracy
* Classification Report
* Confusion Matrix

# 

# Exploratory Data Analysis

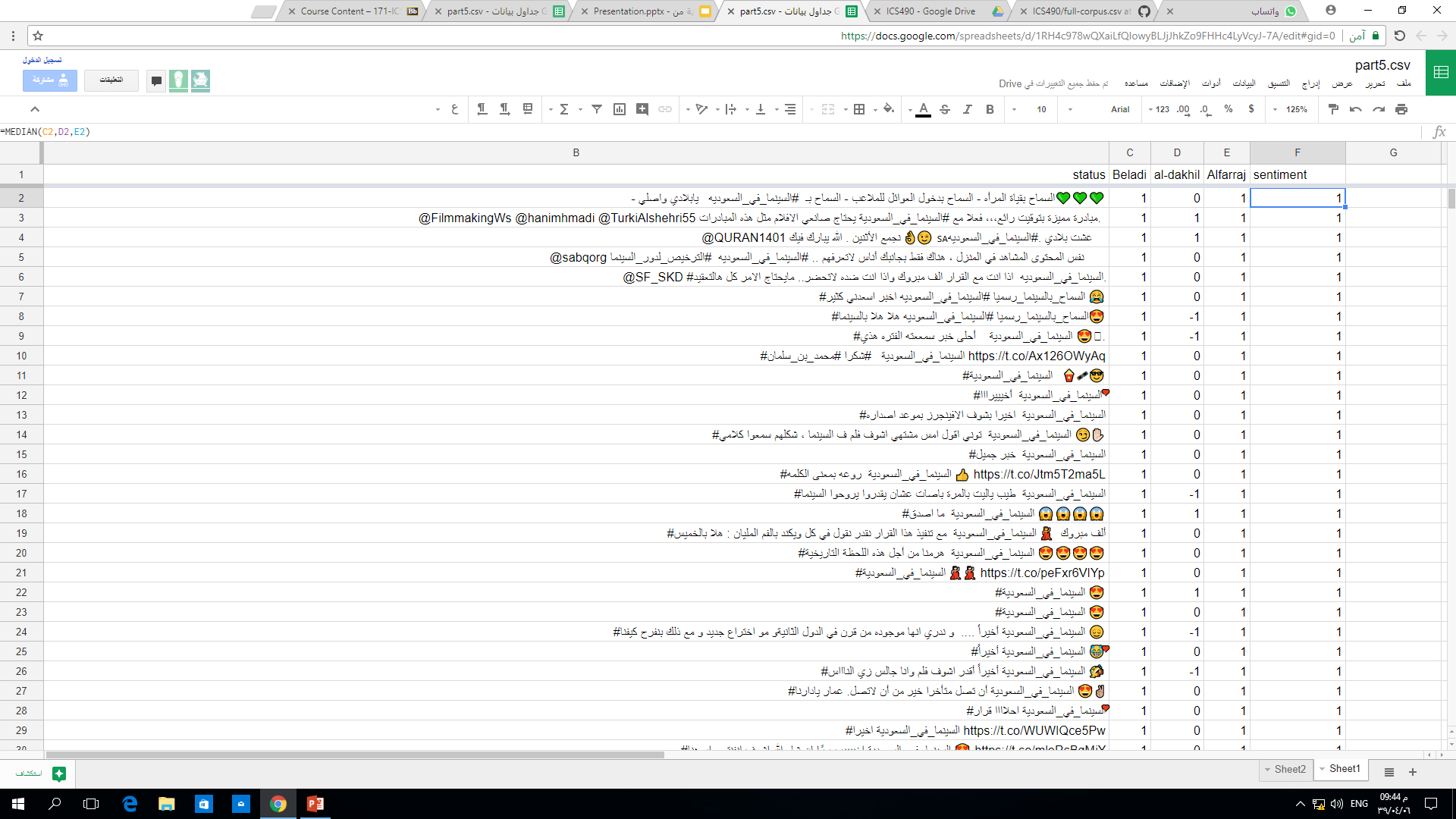
* Number of gathered tweets including retweets: 19548, 2327 of which are real Tweets.
* **Total** dataset: **5328** tweets (2327 tweets + 3001 of tweets are gift from Mustafa’s group)
  + **Positive** tweets : **1458** tweets
  + **Neutral** tweets: **2509** tweets
  + **Negative** tweets: **626** tweets
  + **Spam** tweets: **735** tweets

# Procedure

## Data Gathering

Using twitter streaming API to obtain the tweets containing the hashtag.

## Data Labeling



## 

## 

## Data Cleaning

* Translating numbers from Hindu numbers to Arabic numbers.
* Replacing links, numbers, accounts with tokens.
* Removing the punctuation and arabic diacritics.
* Uniforming multiple characters to their original form.
* Removing Repeated letters.

## 

## Features Extraction

TfidfVectorizer (CountVectorizer followed by TfidfTransformer)

## 

## Classifiers

* Running the pipelines of classifiers with the exported features matrix.
  + Each pipeline contains TfidfVectorizer followed by the classification method.
* Feeding the pipelines to an ensemble voting classifier.

# 

# 

# 

# Performance Results

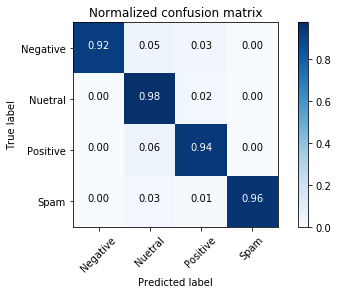
## Support Vector Machines Classifier

Accuracy: 95.87%

Classification report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | Support |
| Negative | 1.00 | 0.92 | 0.96 | 64 |
| Neutral | 0.95 | 0.98 | 0.96 | 258 |
| Positive | 0.94 | 0.94 | 0.94 | 144 |
| Spam | 1.00 | 0.96 | 0.98 | 67 |
| Avg / Total | 0.96 | 0.96 | 0.96 | 533 |

Confusion matrix:



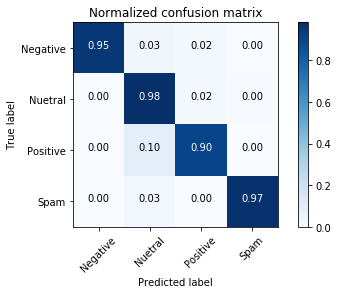
## Multinomial Naïve Bayes Classifier

Accuracy: 95.5%

Classification report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | Support |
| Negative | 1.00 | 0.95 | 0.98 | 64 |
| Neutral | 0.93 | 0.98 | 0.96 | 258 |
| Positive | 0.96 | 0.90 | 0.93 | 144 |
| Spam | 0.98 | 0.97 | 0.98 | 67 |
| Avg / Total | 0.96 | 0.95 | 0.95 | 533 |

Confusion matrix:



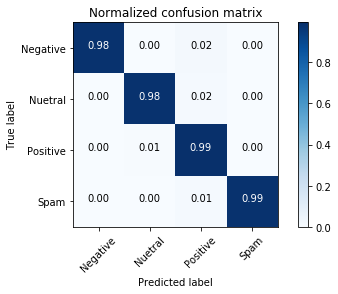
## Random Forests Classifier

Accuracy: 98.69%

Classification report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | Support |
| Negative | 1.00 | 0.98 | 0.99 | 64 |
| Neutral | 1.00 | 0.98 | 0.99 | 258 |
| Positive | 0.96 | 0.99 | 0.98 | 144 |
| Spam | 1.00 | 0.99 | 0.99 | 67 |
| Avg / Total | 0.99 | 0.99 | 0.99 | 533 |

Confusion matrix:



## 

## 

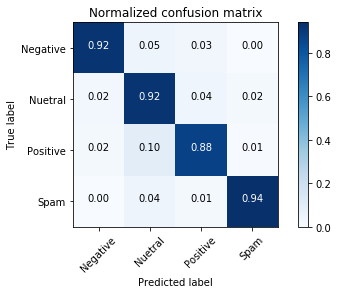
## Logistic Regression Classifier

Accuracy: 90.99%

Classification report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | Support |
| Negative | 0.87 | 0.92 | 0.89 | 64 |
| Neutral | 0.92 | 0.92 | 0.92 | 258 |
| Positive | 0.90 | 0.88 | 0.89 | 144 |
| Spam | 0.93 | 0.94 | 0.93 | 67 |
| Avg / Total | 0.91 | 0.91 | 0.91 | 533 |

Confusion matrix:



## 

## 

## Voting Classifier

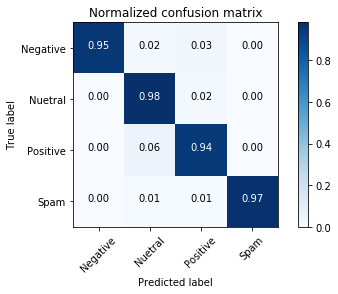
Ensembling the earlier classifiers

Accuracy: 96.62%

Classification report:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-score | Support |
| Negative | 1.00 | 0.95 | 0.98 | 64 |
| Neutral | 0.96 | 0.98 | 0.97 | 258 |
| Positive | 0.94 | 0.94 | 0.94 | 144 |
| Spam | 1.00 | 0.97 | 0.98 | 67 |
| Avg / Total | 0.97 | 0.97 | 0.97 | 533 |

Confusion matrix:



# Parameters

## Support Vector Machines Classifier

The best parameters obtained by GridSearch were:

For the classifier:

* C=1.3
* kernel='linear'
* decision\_function\_shape='ovo'

## Multinomial Naïve Bayes Classifier

The best parameters obtained by GridSearch were:

For TfidfVectorizer:

* max\_df=0.4

For the classifier:

* alpha=0.05

## Random Forests Classifier

The best parameters obtained by GridSearch were:

For the classifier:

* n\_estimators=150

## 

## Logistic Regression Classifier

The best parameters obtained by GridSearch were:

For the classifier:

* class\_weight='balanced'
* C=1.4
* max\_iter=10

## 

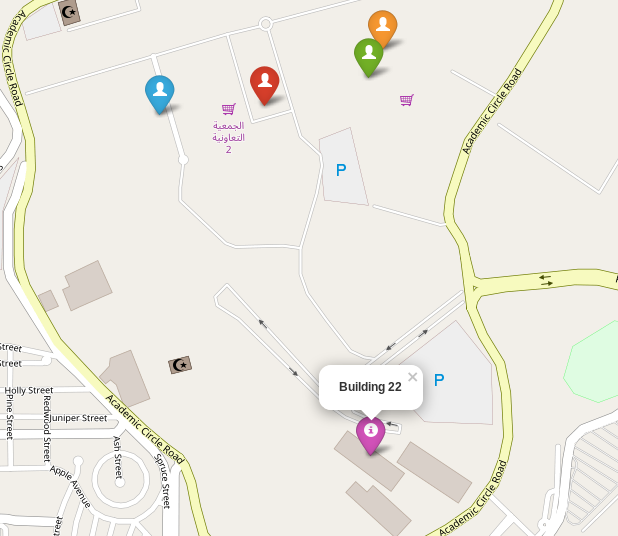
## Voting Classifier

We used the hard voting only.

# 

# Tweets’ Geological Distribution

Unfortunately, none of the 5328 tweets contained the geological coordinates. Hence, there is no geological distribution for the tweets.



# 

# Conclusion

There is no best algorithm for all datasets, but for this dataset the best classifier algorithm is Random Forest Classifier algorithm. However, the number of the tweets isn't significant, 5000 tweets are big to give a statistical significance nor ensures that the Random Forest Classifier is the best for our purpose.