Real time sentiment analysis

Introduction

Sentiment analysis means classifying words to either be negative, positive, or neutral mainly through text analysis and natural language processing,

this benefits the company to gain an insight and overview of what the customers feel about the brand and what should the company do to make the customer feel better

example:

when Cristiano Ronaldo rejoined Manchester united

Cristiano Ronaldo's United shirt breaks all sales records where in the first twelve hours, the shirt generated around 38 million euros.

Companies use sentimental analysis to find out what brings customers to buy the new jersey and taking decisions like high production of the shirt,

Now this requires instant knowledge of the customer needs from the company and so requires feeding live data to machine learning systems (live sentiment analysis) to quickly make decisions.

The data is gathered mainly from social media platforms because people get to express themselves freely, this can be done using web scrapping technologies like selenium and beautiful soup however,

This is not the official method, so twitter constructed an API for legal life data gathering which is twitter API

After getting the credentials by walking through developer's account steps

You will be able to collect the data.

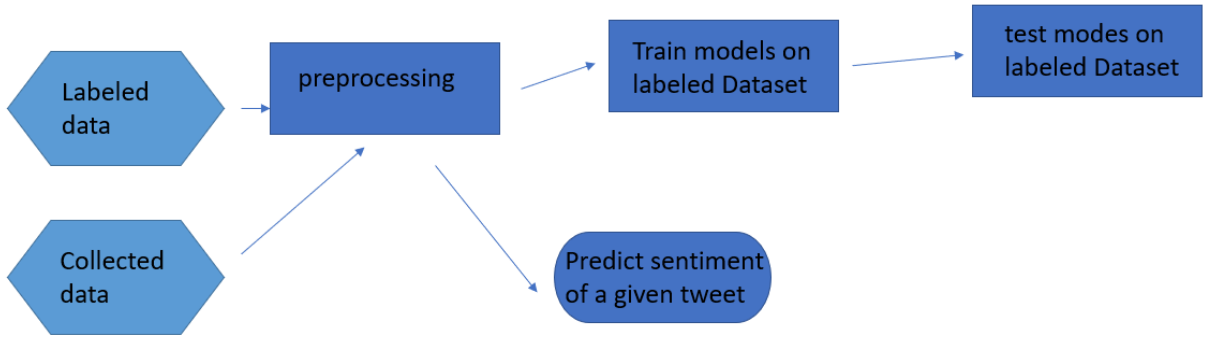


Fig 1: steps of sentiment analysis

Steps:

1-gathering data from twitter:

- Developer's account registration get credentials needed so that API accepts the attempt for data gathering
- Use tweepy (python library) to access the API and stream the Tweets
- Enter the keyword you are searching for

2- preprocessing labeled dataset:

- Tockenizer breaks raw string into words and sentences
- This helps the machine to understand the meaning of words
- Removing (stop words) they are commonly used words like (a) and (is)
- HashingTF to vectorize the features with a standard dimension(262,144)

3- training and testing machine learning models of whole dataset (1600000):

Logisitc regression : one of the most famous classification models used for simplicity and efficiency using grid builder for best estimator result

Naïve bayes : simple yet accurate classifier

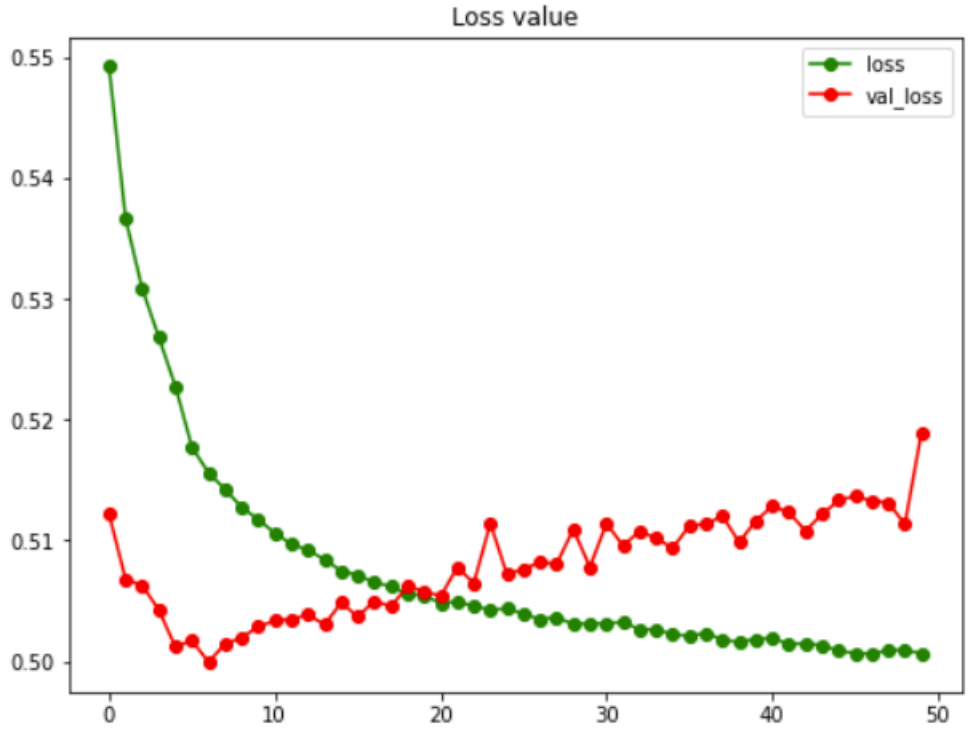
GBT regressor: used in classification and regression building random forest classifier (ensemble and minimize cost function)

FM classifier: Factorization machines that build interaction between features

Neural network:

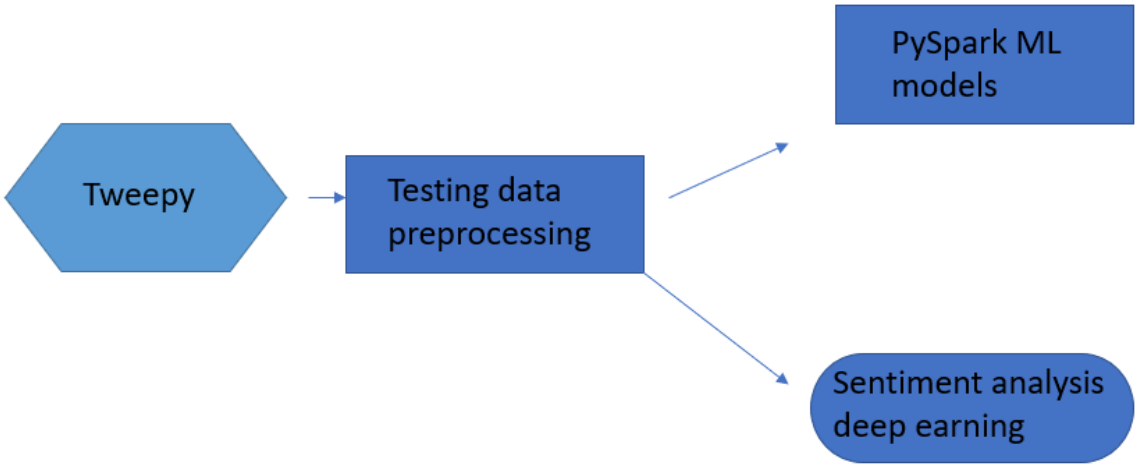
```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 128)               76928
_________________________________________________________________
dense_1 (Dense)              (None, 128)               16512
_________________________________________________________________
dropout (Dropout)            (None, 128)               0
_________________________________________________________________
dense_2 (Dense)              (None, 64)                8256
_________________________________________________________________
dense_3 (Dense)              (None, 12)                780
_________________________________________________________________
dense_4 (Dense)              (None, 8)                 104
_________________________________________________________________
dropout_1 (Dropout)          (None, 8)                 0
_________________________________________________________________
dense_5 (Dense)              (None, 1)                 9
=================================================================
Total params: 102,589
Trainable params: 102,589
Non-trainable params: 0
_________________________________________________________________
```

NN archeticture



Accuracy comparison between models:

Model	Test accuracy
Logistic regression	74
Naïve bayes	76
GBTclassifier	70
FMclassifier	71
Neural network	76



Scripts Flowchart

Multilingual sentiment analysis can be done using

Multilingual Sentiment Classifier for most of languages

pysentimiento A Python multilingual toolkit for Sentiment Analysis and Social NLP tasks (pythonawesome.com)

if it is based on my script it will be simply training the model on both data languages integrated with each other

```
StreamingQueryException: Connection refused: connect
=== Streaming Query ===
Identifier: Ahmed [id = a691bb33-6c7d-4852-9ca3-432f9a19cc35, runId = 71c2f108-d803-4b9b-9c7d-90171c9d25bc]
Current Committed Offsets: {}
Current Available Offsets: {TextSocketV2[host: 0.0.0.0, port: 5555]: -1}

Current State: ACTIVE
Thread State: RUNNABLE

Logical Plan:
Repartition 1, true
+- Project [word#17, polarity#20, subjectivity_detection(word#17) AS subjectivity#24]
   +- Project [word#17, polarity_detection(word#17) AS polarity#20]
      +- Project [regexp_replace(word#15, :, , 1) AS word#17]
         +- Project [regexp_replace(word#13, RT, , 1) AS word#15]
            +- Project [regexp_replace(word#11, #, , 1) AS word#13]
               +- Project [regexp_replace(word#9, @\w+, , 1) AS word#11]
                  +- Project [regexp_replace(word#6, http\S+, , 1) AS word#9]
                     +- Filter AtLeastNNulls(n, word#6)
                        +- Project [CASE WHEN (word#3 = ) THEN cast(null as string) ELSE word#3 END AS word#6]
                           +- Project [word#3]
                              +- Generate explode(split(value#0, t_end, -1)), false, [word#3]
                                 +- StreamingDataSourceV2Relation [value#0], org.apache.spark.sql.execution.streaming.sources.T
extSocketTable$$anon$1@396b89b5, TextSocketV2[host: 0.0.0.0, port: 5555]
```

Error :query.awaitTermination()

References:

https://developer.twitter.com/en

http://spark.apache.org/docs/latest/api/python/

https://www.youtube.com/watch?v=wlnx-7cm4Gg&list=PL5tcWHG-UPH2zBfOz40HSzcGUPAVOOnu1

https://www.youtube.com/watch?v=B-x58mOUEbw

multilingual-sentiment-classifier · PyPI

A Python multilingual toolkit for Sentiment Analysis and Social NLP tasks (pythonawesome.com)