Università degli Studi di Trieste

dipartimento di matematica e geoscienze

Corso di Laurea Magistrale in Data Science and Scientific Computing

Natural Language Processing: Project Report

Student:
Pietro Morichetti - SM3500414

Anno Accademico 2020-2021

A Text Classification Problem: Sentiment Analysis on Pfizer Vaccine's Tweets

This report is written for the Natural Language Processing (NLP) Course at Trieste's University, part of the Master Degree in Data Science & Scientific Computing.

The report will provide Text Sentiment Classification on the Pfizer vaccine, and it takes the Twitter social network platform as source to gather a bunch of tweets on which to apply the sentiment analysis.

The report is divided into five sections:

- a section to introduce the study case;
- a section to explain how the data collection is performed;
- a section to explore the data and find some interesting context information;
- a section to implement several machine learning models.

This specific Sentiment classification problem may be taken on in several ways, in this report it is taken on by using the multi-class classification models and a more sophisticated deep-learning model. The entire project is developed in python, both in local and on Google Collab platform.

Context Introduction

On the 31^{th} of dicember in Wuhan (China) some cases of pneumonia have been observed, days have passed and the cases grew exponentially. At this point governments' attention was focused on this disease that was diffused in the majority of China. On the 31^{th} of January the first cases in Italy are identified, the country soon became one of the epicenters of this infection, a disease that will drastically change the life style of billions of people all around the world.

In spite of the serious situation we are still experiencing today, nations all over the world have started a real race to find an effective solution, and with timely effects: the search for a vaccine.

Definition of the sentiment problem

Fortunately, at this time different drugs are being offered and some of them are approved by the scientific community and dedicated authorities, for example: Moderna, AstraZeneca, Sputnik and Pfizer. However, several voices have expressed their positive or negative opinion on these vaccines, and on different communication channels.

In a digitalized and globalized society like the one we are living in now, more digital solutions exist to express one's own opinion on something. On these platforms everyone can say what they want and everyone can reply: these tools become, at this time, very important to be able to influence the general idea on a specific argument, leading towards considerable effects on ordinary life as well.

From the moment people started to use these tools a lot, it was clear they could be used to do different kinds of analysis, like: collective psychology analysis, business analysis and of course sentiment analysis on public opinions. These analyses provide interesting results on the comprehesion of social phenomena.

This report will refer to the social platform Twitter to analize the public opinion on a specific vaccine developed for the Covid-19: the Pfizer vaccine.

Twitter Data Collection

The choice to use Twitter is justified by the fact that among the collective communication channels, it is the most used and it provides a likely sample of the global population with millions of tweets (i.e. messages) per day. Twitter is often used by influent persons, public and private authorities to share news and information of common interest.

This is very important because, implicitly, Twitter is recognized as "official" communication channel to share news. The results affect the "communnication power" of Twitter's messages and are able to awaken the awareness of a lot of people.

Connection to Twitter API's

To enter Twitter's universe it is necessary to create a Twitter Developer account at the official website. Then, some private keys will be assigned to connect to Twitter's API in order to downlaod data; in particular, the following keys are provided:

- **consumer key & secret consumer key** - think about them as the username and password that represents your Twitter developer app when making API requests.
- **access token & access token secret** - An access token and access token secret are user-specific credentials used to authenticate OAuth 1.0a API requests. They specify the Twitter account the request is made on behalf of.

Note that these keys might have a different name outside of this report; for detailed information, please have a look on Twitter's documentation.

Data mining

Using Twitter is very simple and intuitive, however the real situation behind the front-end is quite complicated, for example the way the messages are structured. Indeed, each posted tweet, each reply or quote is considered like a new "object": unique in its informative content but it is regulated according to precise structural rules. In other words, all the tweets have the same characteristics, but each tweet is unmistakable also because they have a unique ID. By considering its complex structure, it is not so simple to get the needed information.

Beyond the complications caused by the tweet structure, some limitations imposed by Twitter or that could occur during the tweet mining phase should also be considered; here, some of them are presented.

constraints imposed Since Twitter is used by many developers and companies, it is necessary to fix a quote for the requests sent to the servers to avoid critical problems; for this reason, Twitter decides that only a certain number of tweets can be required every 15 minutes. This kind of constraint is valid just for tweets that are not considered as *live-stream*, this means that the tweet brings all its information. Meanwhile, tweets that are in live-stream can be downloaded without quantity restrictions because they are just composed by the text of the message and their ID.

Beyond these limitation, there are others due to privacy: for instance it is not possible to access the geo-coordinates of the users that send the tweet, the only available geographical information is the provenance location of the user, provided by the user itself. Another constraint could derive by the kind of Twitter account, in fact Twitter provides different accounts with different tools to collect data.

random constraints This kind of constraint on data is considered random because it should be available a priori, but for some reason it is not for a while or even forever. These constraints could be the elimination or the suspesion

Parameter Name	**Type**	**Brief Description**
created_at	datetime	date and time in which the tweet is created
id	int	ID of the tweet
user.screen_name	string	username of the tweet owner
user.id	int	ID of the tweet owner
user.verified	bool	if the owner of the tweet verify it's identity
user.location	string	location provided by the owner of the tweet
user.followers_count	int	number of users following the owner of the tweet
retweet_count	int	number of users that have re-tweeted the tweet
favorite_count	int	number of likes obtained by the tweet
hashtags	list of string	hashtags in the tweet
replies	list of string	comments of the tweet
n_replies	int	number of comments of the tweet
full_text	string	text of the tweet

Table 1: Tweet's Inforamtion gathered.

of the message by the user or by Twitter, because it violetes Twitter rules. As for the messages, this kind of limitation might be applied on the user account and its information could not be available anymore.

While implementing a software for extracting information from Twitter, it is good to keep in mind that some of the previous limitations might occur. So, if they are not properly handled, the program could not work very well.

In this first phase on data mining the tweepy python library is used, both to connect the program to the Twitter's API, to collect tweets and to extract the needed information. In particular, 20K tweets are downloaded without considering re-tweets or replies; to execute this operation the following query was used:

'(Pfizer OR Pfizer vaccine) AND -filter:retweets -filter:replies'

Following this stage, the next one is the extraction of specific information from the tweet; in particular, comments were collected but they were not used.

Assignment of sentiment labels

At this point a dataset is created from scratch with different info associated to each tweet, but something is still needed: remember that the target of this paper is to realize a machine learning model to infer the sentiment related to tweets, based on the text of the tweet. For that goal, these models need to be trained by considering given examples. However, the built dataset does not present this kind of information yet.

Different strategies can be used to label the tweets, like assigning lables "by hand" for some and then training the models on them, or training the models on another similar already labelled dataset.

The used strategy is to consider some libraries with pre-trained models and use one of them to label the tweets: the pre-trained model provides a set of percentages for each sentiment (negative, neutral and posite). Different libraries were used, like VADER, TextBlob, but among the possible choices here the Hugging Face package is used.

Even if the Hugging Face package is considered, rest the fact that, in any case, it is not so reccomended to apply a pre-processing on text because may effect the labeling and this is true for most of the sentiment libraries.

So, it could be done a very soft pre-processing cleaning on tweets, by just removing html strings, since they are not necessary. The script also presents a method to get the higher sentiment among the proposed ones.

A possible operation is to tokenize sentences for each tweet, and then compute the average among scores of the same 'type', in this way we should increase the accuracy of the results. However, since Twitter has no real punctuation rules that we may consider to split tweets in sentences, this task could be quite complex and may be misleading for the purposes of this project, and for that reason it will not be done here.

Data storing

In the last part of this phase the dataset is stored to make it available a posteriori, since it is quite expensive to re-built in terms of execution time. However, since the dataset is a complex structure, it is stored in JSON format that has the capacity to mantain the dataset integrity.

Data Exploration Analysis

The actual section is the second phase of the work, i.e. the data exploration both in terms of Natural Language Processing (Topic Modelling, Entity Analysis and PMI), and in terms of descriptive analysis.

In both cases it was preferred to place a preliminare phase before, in which the text of the tweets is preprocessed to reduce their variance. This task is performed by a dedicated function where some cleaning procedure is

Topics	**Related Words**									
T1	jab	wait	fulli	think	go	work	look	better	tomorrow	week
T2	dose	got	shot	second	effect	today	get	feel	arm	day
T3	covid	peopl	receiv	israel	heart	case	india	inflamm	world	examin
T4	moderna	deal	time	know	said	booster	jampj	canada	come	problem
T5	avail	appoint	covid	new	need	moderna	pfizerbiontech	amp	astrazeneca	read

Table 2: Words associated to LDA topics

applied, like: removing digits, transforming emojies, tags and other. Moreover, words are transformed from uppercase to lowercase, stopwords are removed and if it is necessary, stemming is applied. This function is able to return a list of the cleaned tweets, or a tokenized list version of the tweets.

Data NLP analysis

As introduced before, an NLP analysis is shown here, in particular a Topic Modelling analysis, an Entity analisys and an exploration of the PMI are performed.

Topic Modelling Why is a topic modelling analysis required? The reason is related to the data, since the tweets dataset is obtained directly from Twitter, there is no certainty that most of them are associated with the Pfizer topics. So, before everything it is necessary to be sure about that and that these steps are followed:

1. exec an hard pre-processing on twees;
2. construct a dictionary on the remained words, excluding the less and most frequents;
3. define the LDA's corpus by means of the word2vec model;
4. define the TF-IDF rapresentation of the LDA corpus;
5. train the LDA model to infer on the distribution of 5 topics (number choosen by default).

In conclusion the following table 2 is obtained.

Note some terms like: dose, covid, pfizierbiotech, india and inflammation; so, it is correct to say that collected tweets are sufficiently related to the main argument. However, note that Topic Modelling is a type of stochastic analysis, then each iteration might lead to different results; moreover, the pre-processing phase and the parameters definition of the model influence the performace of the obtained results.

Entity Analysis The second analysis applied is the research of entities in the tweets that could help to better understand what people are more focused on; the procedure is shown below.

1. apply a pre-processing on tweets to reduce the variance;
2. determine the entity of each token by means of NLP libraries;
3. determine the more frequent entities.

The identification of the entity shows that the most frequent entities are: GPE, ORG and PERSON as shown by this picture 1.

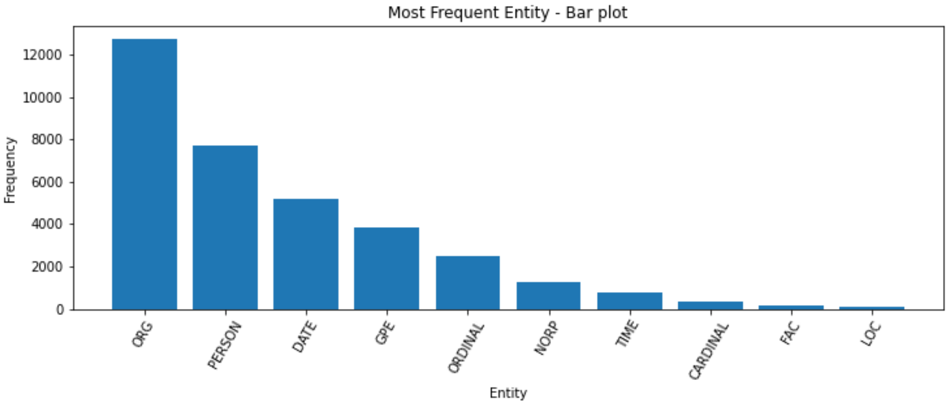

Figure 1: Most frequent entity encountered in the tweets.

Moreover, it might be of interest to understand how the sentiment of a tweet, where there is a certain entity, changes in time like shown in figure 2.

As shown by the picture, if the 'pfizer' entity is considered, its positive sentiment trend is mostly dominant to the other sentiments during the considered period of time.

PMI Another useful analysis is to search for terms that are most likely to be found together; so, the analysis try to define the collocations set, as in table 3.

The results of the analysis show that the most part of Twitter's users speak about topics related to Covid-19 and vaccines effects. This shows how people are quite worried about the negative effects that could appear from them.

This phase reveals interesting information, however it is good to note that there is room for improvement. Indeed, Topic Modelling executed does not present any tuning phase for the hype-parameters, like the number of topics.

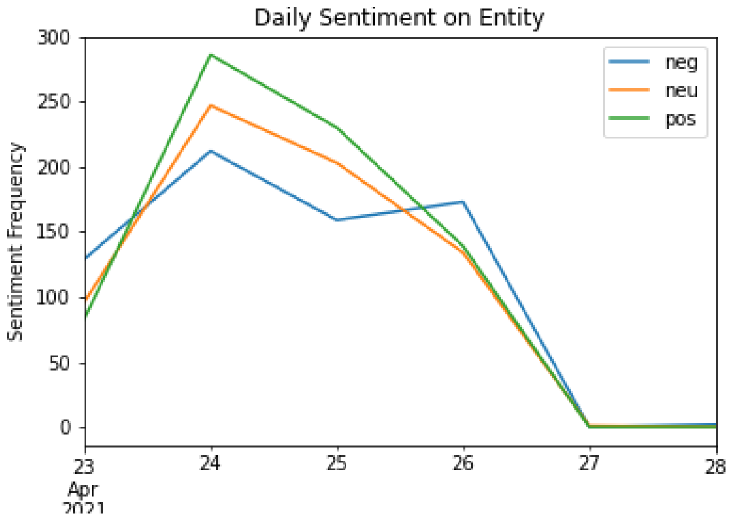

Figure 2: Sentiment trends associated to the 'pfizer' entity.

Instead, in the entity analysis, the associated sentiment is based exclusively on the general sentiment of the tweet: this means that the sentiment of the tweet might not be referred to the considered entity, but to another entity present in the tweet. A way to improve the association among tweet sentiment and entity could be to execute a preliminary analysis on the semantics of the tweet.

Data descriptive analysis

A very frequent and important type of analysis is descriptive analysis, where one tries to extract behaviours and aspects of the data that might not be visible at first glance. It is possible to execute a lot of explorative analysis on data, in this paper some of them are shown.

Most frequent terms One might be interested in understanding what the most frequently used words in the Pfizer vaccine and Covid-19 context are, the following picture shows that terms such as pfizer, vaccine, second and dose have a high level of frequency in the order of millions.

A more fun representation of the most used terms might be provided by means of a Word Cloud as shown by the following picture 3.

Or again it might be interesting to understand what the most frequent hashtags are, since they can assume great significance in social context, as shown by the following picture 4.

Terms	Score
heart_inflamm	331.863
raw_materi	238.647
israel_examin	198.439
examin_heart	182.204
inflamm_case	165.408
pfizer_vaccin	142.228
neurolog_degen	137.768
covid_vaccin	116.6277
caus_neurodegen	112.576
al_neurolog	109.096
alzheim_al	106.588
sore_arm	104.236
second_dose	100.465
trigger_alzheim	97.504
bee_gee	85.0
pfizer_shot	84.327
neurodegen_diseas	82.586
gee_singer	76.246
barri_gibb	73.751
singer_barri	71.579

Table 3: PMI table, collections are mostly refered to the effects of the Covid-19 and vaccines.

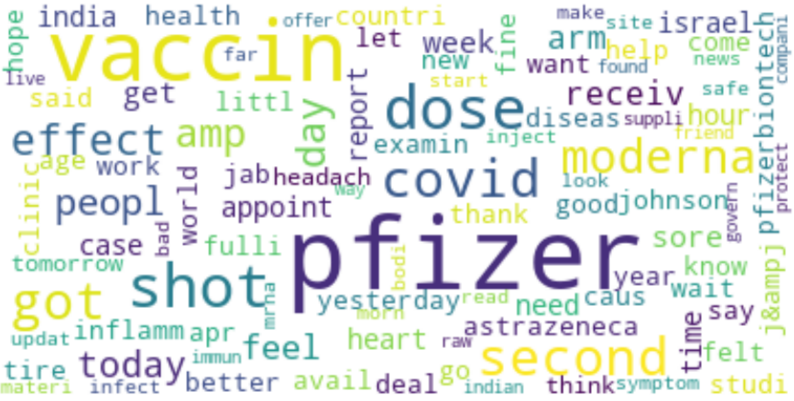

Figure 3: Word Cloud of the most frequent words.

Investigation of users Beside the term analysis, one might be interested in determining hidden users information: for example the location of the users, as shown by the picture.

In particular, the majority of users do not provide this kind of information. However, despite the fact that some of the users have provided it, it is good to take into account that location is not a reliable information. In fact, users can insert any place as location (even invented) and that makes the location not so relevant, or even leading to wrong conclusions.

Further information on Twitter users is checking if they have verified their idenitity, the following picture 5 shows that the majority of them tend to not verify their own account.

One might be interested in understanding which users are more active, based on the number of tweets posted: this is something important because active users, that have a lot of followers, might be able to influence the

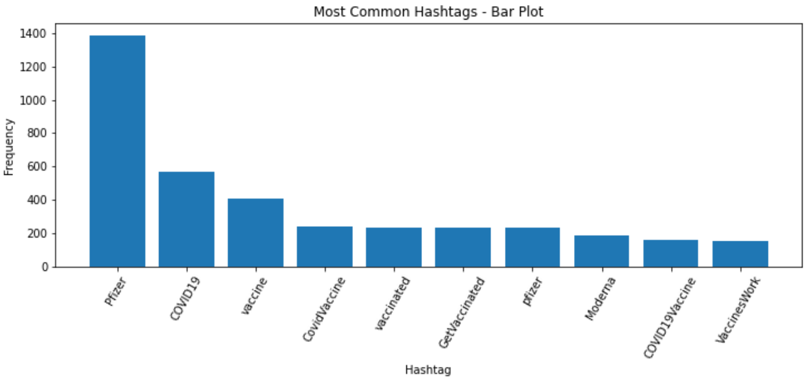

Figure 4: Bar plot of the most common hashtags.

opinion of other users, as shown in the picture 6.

Overview on tweets sentiment As last descriptive analysis a representation of the general feeling over Pfizer is shown by the following picture 7.

In particular, even if the majority of the tweets have a neutral sentiment, the dataset is balanced enough.

Models Implementation

In this section some machine learning models used to perform the Text Classification on the tweets sentiment on the Pfizer's vaccine are presented. In particular, two families of classifiers: multi-class classification and neural networks.

Classifiers are evaluated according to some metrics as for example: precision, recall, F1_score, etc.. Three methods to show some of these measures were defined, one to plot the confusion matrix, and other two for ROC and Precision-Recall Curve.

Here is the definition for these metrics:

- **Precision -** represents the accuracy with which the model is able to predict the positive class, it is given as the ratio between TP and TP + FP

- **Recall -** represents the rate of instances of the positive class that are correctly recognised by the model, it is given as the ratio of TP to TP + FN.

- **F1 Score -** represents the combination of Precision and Recall, in particular it constitutes the harmonic mean between these two metrics. It

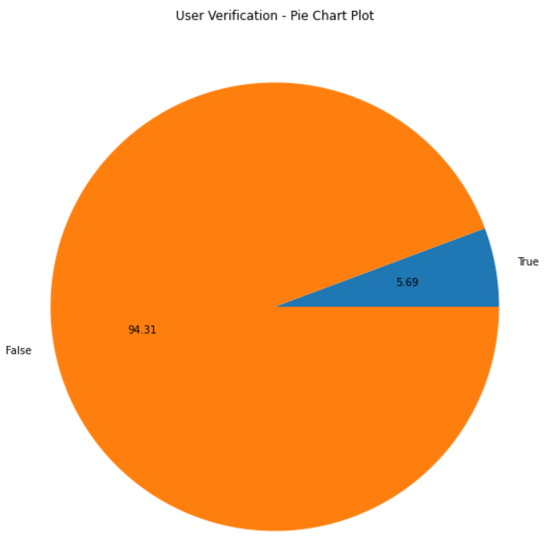

Figure 5: Pie chart for user verification.

tells you how precise your classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances).

- **Macro Avg -** it can be used when you want to know how the system performs overall across the sets of data.
- **Micro Avg -** when there is a need to weight each instance or prediction equally.
- **Confusion Matrix -** is a way of representing the effectiveness of a model by reporting the number of correctly, or incorrectly, classified instances in a matrix.
- **ROC Curve -** is a metric for assessing the performance of a classification model.
- **PR Curve -** similar to the ROC Curve, with the difference that it is a preferred metric for unbalanced datasets.

In the following section will the evaluation of the models according to these metrics be presented.

Multi-Class Classification Model

Multiclass classification is a classification task with more than two classes: each sample can only be labeled as one class. That is exactly as in our

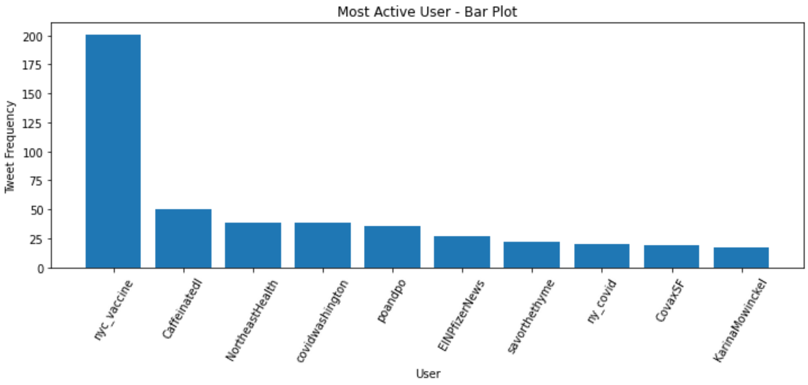

Figure 6: Bar plot of the most active user according number of tweets posted.

	Regularized			Regularized + Feature Selection		
	Precision	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
Class 0	0.66	0.68	0.67	0.67	0.64	0.65
Class 1	0.73	0.64	0.68	0.69	0.68	0.69
Class 2	0.64	0.74	0.68	0.63	0.69	0.66
Accuracy	-	-	0.68	-	-	0.67
Macro avg	0.67	0.69	0.68	0.67	0.67	0.67
Micro avg	0.68	0.68	0.68	0.67	0.67	0.67

Table 4: Evaluation table for Logistic Regression model in case of regularization, and regularization + feature selection

case, since we have sentiment feature that can take one of the three different values: positive, negative or neutral.

In the following subsection we are going to try some strategies: LogisticRegresssion, OnevsAll and OnevsOne.

Logistic Regression Before training the model a cross-validation phase is performed to determine the appropriate value for the regularization parameter, used to set how strong should the regularization be. After that, the model is trained on the train set and then tested on the test set. To try to improve the model, a further analysis is made on the data: try to consider the most related features with the output target class. The following table shows the meausures for the considered metrics: class 0 and class 2 have a lower precision means that less relevant items are selected in respect of the entire provided. Moreover, both configuration has basically the same relevant measures.

Moreover, the following picture 8 shows the confusion matrix, the ROC curve and the PR curve, as it is showen a good results are achived.

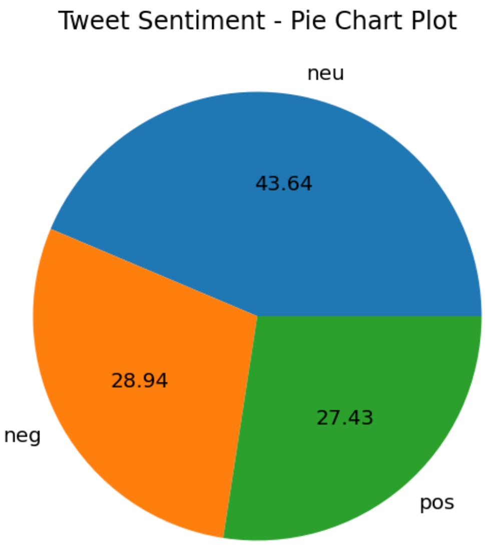

Figure 7: Pie chart of the general sentiment of the users on the Pfizer topic.

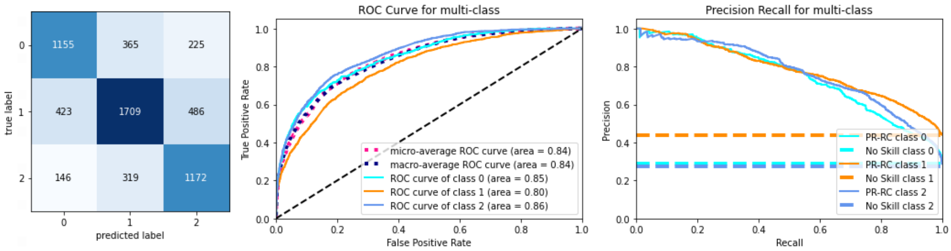

Figure 8: Logistic Regression Model 2° configuration, from left to right: confusion matrix, ROC curve and PR curve

OnevsAll classifier OnevsAll is a kind of strategy used in case of multi class classification: one class at a time is choosen as positive class, while the others are grouped as the negative class. So, it leads to a binary classification and, in this specific case, the classifier is a LINEARSVC, similar to the Support Vector Machine.

Note the following indices table and 9 with the confusion matrix, ROC curve and PR curve.

OnevsOne classifier OnevsOne is a kind of strategy used in case of multi class classification: classes are considered in pair by excluding the others, so a number of classifiers equal to $\frac{n\_classes \cdot (n\_classes - 1)}{2}$ are examined. Also in this case, this strategy leads to a binary classification and again with a LINEARSVC as classifier.

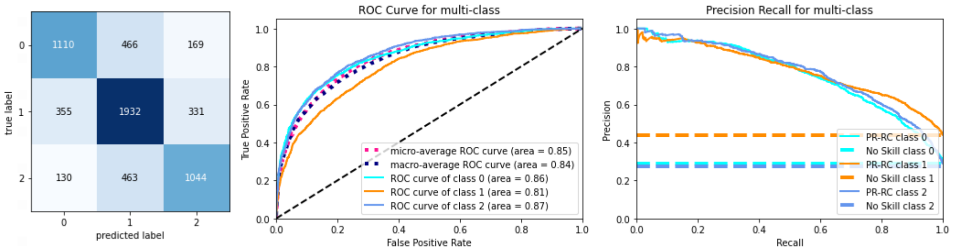

Figure 9: OnevsAll, from left to right: confusion matrix, ROC curve and PR curve

	One vs All			One vs One		
	Precision	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
Class 0	0.70	0.64	0.66	0.70	0.63	0.66
Class 1	0.68	0.74	0.71	0.68	0.74	0.71
Class 2	0.68	0.64	0.66	0.68	0.65	0.66
Accuracy	-	-	0.68	-	-	0.68
Macro avg	0.68	0.67	0.68	0.69	0.67	0.68
Micro avg	0.68	0.68	0.68	0.68	0.68	0.68

Table 5: Evaluation table for OnevsAll and OnevsOne configuration models.

Note the following indices table and picture 10 with the confusion matrix, ROC curve and PR curve.

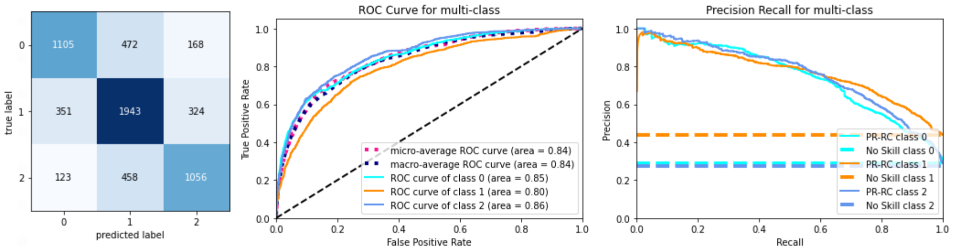

Figure 10: OnevsOne, from left to right: confusion matrix, ROC curve and PR curve

As it is shown there is no clear difference among the two strategies since they have the same measures.

Neural Network

In addition to purely mathematical models, Deep Learning models are also used. Here it is proposed one of these models to infer whether a tweet belongs to one of the sentiment classes: CNN.

	Precision	**Recall**	**F1 Score**
Class 0	0.65	0.69	0.67
Class 1	0.68	0.66	0.67
Class 2	0.63	0.62	0.63
Accuracy	-	-	0.66
Macro avg	0.65	0.66	0.66
Micro avg	0.66	0.66	0.66

Table 6: Evaluation table for the CNN model.

CNN Model The CNN architecture requires, in sequence, the following layers: embedding, 1D convolution, activation, 1D global max pooling, dropout, dense, dropout and again dense layer; as optimizer is used Adam and a categorical cross entropy function as loss function. Moreover, the model is trained with 30 epochs and with some weights on the classes in order to give more relevance to the less frequent classes. The performances of the model are shown in the two pictures 11 that report the trend of the accuracy and of the loss.

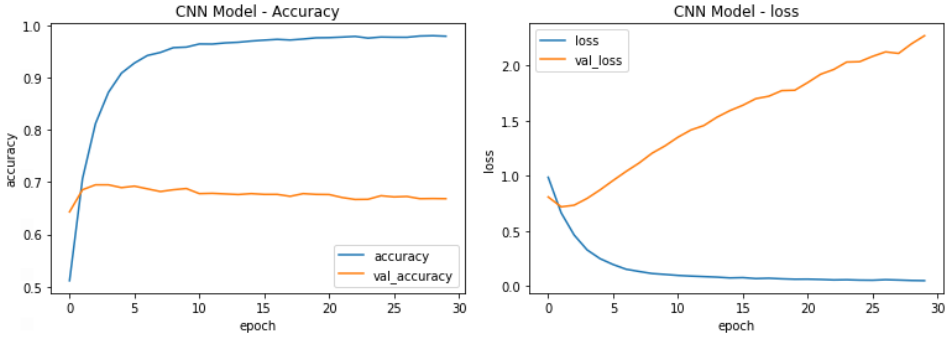

Figure 11: CNN Model, from left to right: plot of the accuracy and plot of the loss

The accuracy of the CNN model tends to converge at one while the validation accuracy, in reverse trend, became worst after each epochs. Also on the CNN - loss plot it is clear that the validation loss tends to increase linearly, while the loss tends to converge to zero.

As in the previous cases, here are the measures for the considered metrics reported.

Moreover, the following picture 12 shows the confusion matrix, the ROC curve and PR curve.

Also in the CNN case, the model provide good results and quite similar to the previous models.

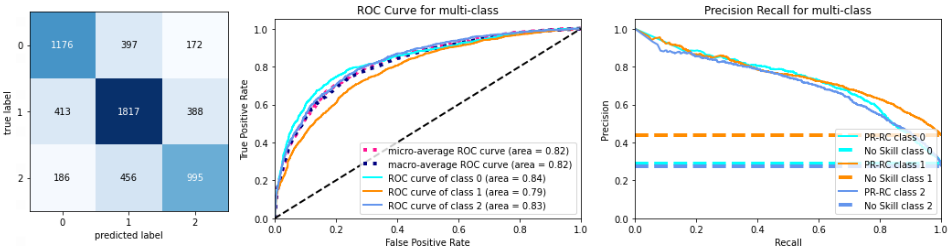

Figure 12: CNN Model, from left to right: confusion matrix, ROC curve and PR curve

Conclusions

This report shows how to apply a Sentiment Analysis on Pfizer's tweets. In particular, it shows a valid procedure to download tweets direcly from Twitter taking in consideration some of the information available. Moreover, two types of analysis on data were offered: NLP analysis and descriptive analysis. In the first case it is shoen that the collected tweets are related to the main topic of this report, which entities are present and what are the most frequent pairs of words; while in the second case, useful information was extracted to better understand the context.

Finally, different machine learning models were proposed to make Text Classification and used to infer the sentiment associated to the tweets. In particular, Logistic Regression, OnevsAll, OnevsOne and CNN models are used with interesting results during the evaluation phase: all of them have provide good results on a real case problem with, for example, an accuracy around 0.68.

All these models have shown similar performance, however it is good to note that there is a margin for improvement as for example trying to tune the parameters, or changing the architecture in case of deep learning models. A further improvement could be done on the pre-processing phase: applying a more focused pre-processing to save some relevant aspects. Additional analysis can be done by applying other models like RNN, Acceptor or BERT to broaden the spectrum of models tested.