parti da definire tecnica ML da utilizzare

1. **Description of the problem (motivation for using ML for this problem)**

As a consulting firm, we have been hired by a renowned German international discount retailer chain to back their upcoming 2022 international advertising campaign aimed at launching a brand new line of fresh food (i.e. food not been [preserved](https://en.wikipedia.org/wiki/Food_preservation) nor [spoiled](https://en.wikipedia.org/wiki/Food_spoilage) yet, as grocery items) via its official Twitter account. The company plans are to start publishing 1 tweet per week starting in April till year’s end, in English language, and focussing on one/more of its existing or newly available food products, present across all its stores. To this end, we have been asked in first place to perform a study meant to understand the main drivers that make certain tweets about food ‘viral’ (i.e. spreading widely and quickly); secondly, the insights gained in the previous step will be used to implement a prediction tool to check each new tweet related to the ongoing campaign before its publication, in order to try to maximize the spreading/impact on the general public. In fact, we have also been charged of the implementation phase of this strategy, so we will have to craft all the weekly tweets tied to this campaign, following the insights of our research: our compensation arrangement is made by a fixed amount plus a variable pay which is based on the count of how many among the tweets we compose will become viral, so to align our interests with the goals of the company. As per our binding contract, given that there are no hard rules defining what a ‘viral’ tweet is (e.g. the number of replies, likes or re-tweets), it will specifically mean a tweet which has been re-tweeted at least 100.000 times.

1. **Problem Statement, and description of proposed solution**

Since often ‘viral’ brings also a time connotation (i.e. a quick spread), we warn the reader that we are not interested in this time dimension of the spread as, following our contract, the aim is to maximize our revenues bringing ideally all the 39 tweets we will compose for the campaign within the ‘popularity target’, regardless of when. We will then use the terms ‘viral’ and ‘popular’ as synonyms in this note.

Binary classification using RF /Logistic Regression/ Decision Trees/ SVM/ BAYES?

1. **Dataset extraction**

We created our reference dataset by randomly scraping all the live stream of tweets older than a week over the 2020 year older than a week via tweepy library and applying 2 filters: the first on the content which must include at least 1 food-related term (we used a dictionary sample to this aim), the second on the number of retweets: to have a balanced sample we want to bi-partiton the space of our tweets into ‘viral’ and ‘no\_viral’ elements with same weight.

As the retweet count and favorite count are always 0 because we use the live streaming API and, as a result, we're scraping the tweets as they are tweeted. At this point, all the tweets have retweet count 0 and favorite count 0 since they were literally just posted! That is, unless the tweet posted is actually a retweet...

To solve the problem of brand new tweets, we used retweets to get the original tweet. This also ensures that our model isn't thrown off when someone with a huge follower count retweets something. Finally, we made sure not to consider the same tweet text twice.

The text of the Tweet and some entity fields are considered for matches. Specifically, the text attribute of the Tweet, expanded\_url and display\_url for links and media, text for hashtags, and screen\_name for user mentions are checked for matches.

Given that non-viral tweets outnumber by far the viral ones, we first count the viral tweets passing all our filters and only then we collect an equal number of ‘no\_viral’ ones, so to create a controlled experiment.

This dataset has then been split into train and test set, and on the former we have applied a k-fold cross validation, to make the estimation of our model more robust.

A lot of factors can affect engagement for a verified user’s Tweet — and they may be completely different from the things that affect Tweets coming from other user groups. As a starting point, we looked for Tweets with specific, measurable features. These are the “hard features” of Twitter: **photos, hashtags, links, videos, tweets containing a number or a digit**.

We analyzed the content of over 2 million Tweets sent by thousands of verified users across different fields over the course of a month. We looked at the numbers of Retweets in that dataset with the specific features mentioned above, and compared that to the average numbers of Retweets for that collection of accounts during that period. In other words, the baseline we compared to was the numbers of Tweets that each user would expect to get anyway. This helps us identify which features within a Tweet have the greatest impact on average Retweets.

The fact is, people don’t engage equally with every Tweet. But now we can confirm that adding video, links and photos all result in an impressive boost in the number of Retweets.

This chart shows the results of a selection of Tweet features, which vary by each industry. Because we are looking at verified accounts that typically have thousands of followers, just about every Tweet results in Retweets, but some see higher levels of retweeting.

1. **Description of indices to measure the quality of the solution provided & experimental procedure for measuring effectiveness**

<https://en.wikipedia.org/wiki/Evaluation_of_binary_classifiers>

|  |  |  |
| --- | --- | --- |
| **Assigned**  **Actual** | **Test outcome *positive*** | **Test outcome *negative*** |
| **Condition positive** | True *positive* | False *negative* |
| **Condition negative** | False *positive* | True *negative* |

1. **Conclusions - Discussion of results of the application of the model**
2. **Future developments and improvements**
3. **References**

* Jenders, M., Kasneci, G., & Naumann, F. (2013, May). Analyzing and predicting viral tweets. In *Proceedings of the 22nd international conference on world wide web* (pp. 657-664).
* Fiok, K., Karwowski, W., Gutierrez, E., & Ahram, T. (2020). Predicting the volume of response to tweets posted by a single Twitter account. *Symmetry*, *12*(6), 1054.
* <http://belindazeng.github.io/goingviral/>
* <https://towardsdatascience.com/using-data-science-to-predict-viral-tweets-615b0acc2e1e>
* <https://towardsdatascience.com/extracting-data-from-twitter-using-python-5ab67bff553a>
* <https://towardsdatascience.com/how-to-scrape-tweets-from-twitter-59287e20f0f1>
* <https://www.sciencedirect.com/science/article/abs/pii/S0957417420306096>
* <https://blog.twitter.com/official/en_us/a/2014/what-fuels-a-tweets-engagement.html>