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. Company’s plans are to publish 1 tweet per week starting in April till year’s end, in English 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, the virality will be measured in terms of the number of single retweets received by an original tweet (i.e. not retweeted). Moreover, given that the retailer chain’s most popular tweet to date counts 350 retweets and there are no hard rules defining what a ‘viral’ tweet is, for us here will specifically mean an original English-language tweet which has been retweeted at least 1.000 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.

This is a supervised learning problem which falls into the Binary classification domain, for this reason we are going to resort to 2 techniques such as RF /Logistic Regression/ Decision Trees/ SVM/ BAYES We will use as predictors the following features:

**time, tweet\_lenght, hashtag\_number, replies\_number, likes\_number**.

At a first sight the number of replies and likes may seem two variables outside our control, nevertheless our client insisted that should we find both significant, they could leverage some questionable likes/replies-boosting technique (e.g. bots, paid likes etc..) for the campaign.

mentions, links, the text length, the

amount of followers, the number of friends, the amount of lists to which

the user is enrolled, a 50-dim image vector, and word features

1. **Dataset extraction**

We created our reference dataset by scraping all the stream of tweets from 21/03/2006 (Twitter inception) to date using Python’s twint library and applying 3 filters. The first is that some entity fields (tweet text, expanded/display urls for links and media, text for hashtags, and screen name mentions) must match at least one food-related keyword (we scraped an English dictionary to this purpose). The second is on the language of the tweet, which must be English. Last filter is on the minimum number of retweets: in fact, to form a balanced sample we want to bi-partiton the space of our tweets into ‘viral’ and ‘non\_viral’ elements with same weight. So, as non-viral tweets outnumber by far the viral ones, after a first download for the popular ones (≥1000 retweets), we counted them and proceeded with a second run where we removed the retweet limit, adding instead a max download limit for the number of tweets, set equal to the previous count.

Finally, we made sure not to consider the same tweet twice, thus removing duplicates (by tweet id).

The final dataset counts 129Kx2 (1GB) tweets, of which almost 50% are viral and 50% are non-viral.

We randomly split the data into train and test set, and on the former we have applied a k-fold cross validation, to make the estimation of our models 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**

Raise awareness

1. **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>
* <https://www.learndatasci.com/glossary/binary-classification/>