parti da definire tecnica ML da utilizzare

1. **Description and statement of the problem**

As a consulting firm, we have been hired by a renowned German international discountretailer chain to back their upcoming 2022 international advertising campaign aimed atlaunching a brand new line of fresh food (i.e. 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 focusing on one/more of its existing or newly available food products, present acrossall its stores. To this end, we have been asked in first place to perform a study meant tounderstand the main drivers that make certain tweets about food ‘viral’ (i.e. spreadingwidely and quickly); secondly, the insights gained in the previous step will be used toimplement a prediction tool to check each new tweet related to the ongoing campaign beforeits publication, in order to try to maximize the spreading/impact on the general public. Infact, we have also been charged of the implementation phase of this strategy, so we will haveto 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 basedon the count of how many tweets -among those composed- will become viral, so to alignour interests with the goals of the company. As per our binding contract, the virality willbe 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 counts350 retweets and there are no hard rules defining what a ‘viral’ tweet is, for us here willspecifically mean an original English-language tweet which has been retweeted at least 1.000times.Since often ‘viral’ brings also a time connotation (i.e. a quick spread), we warn the readerthat 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 forthe campaign within the ‘popularity target’, regardless of when. We will then use the terms‘viral’ and ‘popular’ as synonyms in this note. We also will model the virality of a tweet asa function of specific tweet features[1], excluding the features tied to tweet’s author –which in our case- are outside our control.

1. **Dataset extraction**

We created our reference dataset by scraping all the stream of tweets from Twitter’s in-ception (2006) to date, using Python’s twint library and applying 3 filters. The first filtercontrols for the presence of at least one food-related keyword in some entity fields (tweettext, expanded/display urls for links and media, text for hashtags and screen name men-tions); we scraped an English dictionary for this purpose. The second is on the language ofthe 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-partition the space of our tweets into ‘viral’ and‘non-viral’ elements with equal weight. So, as non-viral tweets outnumber by far the viralones, after a first download of the popular ones (≥1000 retweets), we counted them andproceeded with a second run without min-retweets limit, adding instead a max-downloadlimit for the number of tweets, set equal to the count of viral tweets previously collected.Eventually, not to consider the same tweet twice, we removed duplicates (by tweet id).The final dataset counts 258.730 tweets, balanced between viral and non-viral (50%-50%).Following an exploratory data analysis, we performed a data cleansing dropping some vari-ables which were loosely related to the purpose of our study (e.g.username, place, thumb-nail, etc..), a missing-data imputation was also made specifically for variables which wedecided we will encode as dummy variables (NaN→0). After this first screening, feature en-gineering followed, including transformations such as: response variable from numerical tobinary, with a cutoff set at retweetsΓ1000(retweets→viral), each tweet’s text into its charac-ter count (tweet→tweetlength), the publication date into its weekday (date→weekday),the publication time into its hour (time→hour), the counting of the occurrences of someitems (mentions, urls, photos, hashtags→[...]count), the “dummification” of two-classcategorical variables (quoteurl, video).Once completed variable transformation, we randomly split the dataset into train and testchunks, following a 70/30 proportion. Next, we applied standard scaling feature-wise

1. **Experimental evaluation**

This is a supervised learning problem which falls into the Binary Classification (BC) domain(y∈{0,1}, where 1 for us means “viral” and 0 is “not viral”). Given this kind of problem,we start our analysis employing six techniques commonly known such as: Random For-est (RF), Logistic Regression (LR), Decision Trees (DT), Support-Vector Machine (SVM),Naive Bayes (NB), k-Nearest Neighbors (kNN). All the techniques are implemented usingthe Python library Scikit-learn.

Zero Rule or ZeroR is the benchmark procedure for classification algorithms

**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/>