parti da concordare 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 their publication, in order to try to maximize their 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, with 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 of 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 as been re-tweeted at least 100.000 times.

1. **Problem Statement, a formal statement of the problem**

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.

We created our reference dataset by scraping all the tweets over the 2020 year, and applied 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. 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.

1. **Description of proposed solution**
2. **Description of indices to measure the quality of the solution provided**
3. **Experimental procedure for measuring effectiveness**
4. **Discussion of results of the application of the model**
5. **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://www.sciencedirect.com/science/article/abs/pii/S0957417420306096