Sentiment prediction of Twitter contents

Edoardo Chiò
Politecnico di Torino
Student id: s301486
s301486@studenti.polito.it

***Abstract*—The abstract goes here. Keep it short (approx. 3-4 sentences)**

I. Problem overview

This project concerns a classification problem applied to a collection of Twitter posts (i.e., *tweets*) written by different users. The goal of the project is to perform a sentiment analysis of the posts contained in the dataset.
Training and validation of the models are conducted on a development set, containing 224,994 labelled recordings, while the set to test the models contains 74,999 recordings.
The development set is composed of six fields:

- *sentiment*: sentiment labels;
- *ids*: numerical identifier of the tweet;
- *date*: publication date of the tweet;
- *flag*: query used to collect the tweet;
- *user*: name of the user that posted the tweet;
- *text*: text of the tweet.

The *sentiment* field contains a label for each record. There are two classes: the text is considered having a positive trait if the label value is **1**, instead it is considered negative if the label value is **0**. The classes are not well balanced, indeed there are 130,157 data points having label 1, and just 94,837 having value 0. An exploratory analysis of the dataset was performed, to study the contents of all the fields and understand which features should be taken into account.
The *ids* field (i.e., the row identifiers) presents 278 pairs of duplicates, each one having one row of the related pair of rows associated to a positive sentiment, and the other one to a negative sentiment; since it is not possible to know the correct sentiment for these records, they do not add any insight and so they were removed during the data cleaning. The *date* field shows that the tweets, both in the development dataset and in the evaluation dataset, were posted between April 6th 2009 and June 25th 2009; this suggests that the data should be homogeneous across the datasets, with a similar distribution of sentiment. The *flag* field presents a unique value in all tweets in both the datasets, so it is not useful for the analysis and it can be discarded. The *user* field contains the names of 10,647 distinct users; each user wrote at least 15 tweets (considering the datasets), but some users were much more prolific than others (figure 1). Analysing the user average sentiment distribution (figure 2), it is possible to notice that there are many users whose tweets are almost exclusively positive. This odd behaviour led to a deeper investigation of the texts written by each user; a mean cosine similarity among the tweets posted by the same user was computed (figure 3), and it showed that a small but significant group of users

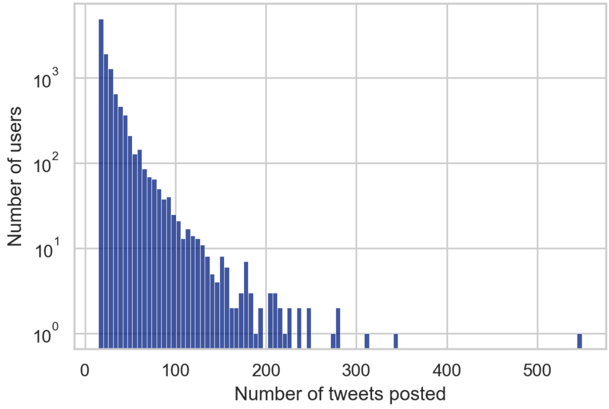

Fig. 1. Histogram representing the number of users per number of tweets

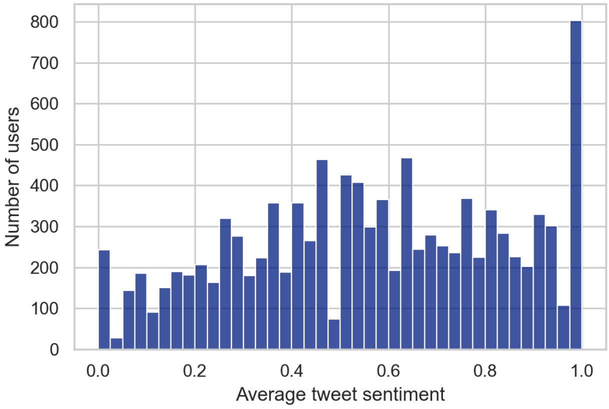

Fig. 2. Histogram representing the number of users per average sentiment

II. Proposed approach

In this section, you will present your solution. Please fill in accordingly.

You can use citations as follows: [1] (you can add BibTeX citations in the *bibliography.bib* file).

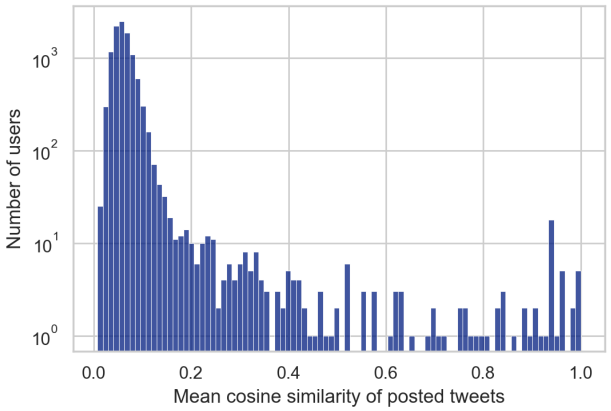

Fig. 3. Histogram representing the number of users per mean cosine similarity computed on the posted tweets of each user

A. Preprocessing

B. Model selection

C. Hyperparameters tuning

III. Results

Here you will present your results (models & configurations selected, performance achieved)

IV. Discussion

Any relevant discussion goes here.

References

[1] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*, vol. 1. MIT press Cambridge, 2016.