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

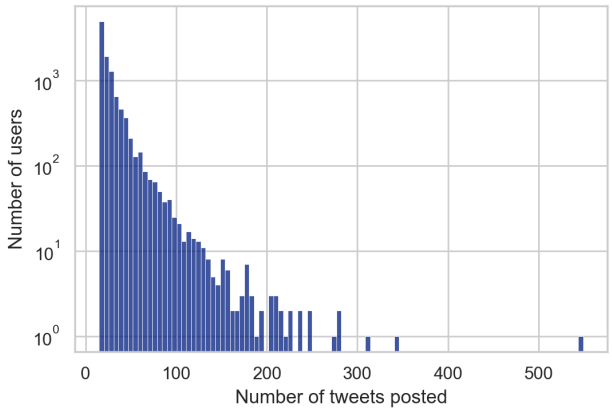

Fig. 1. Histogram representing the number of users per number of tweets

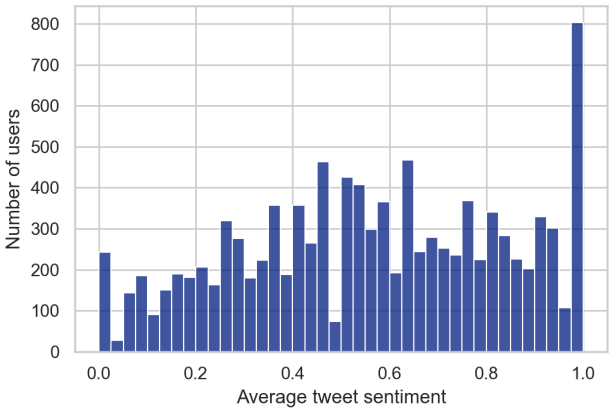

Fig. 2. Histogram representing the number of users per average sentiment

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 used the exact same words in multiple tweets. Some of the users that posted the largest number of tweets, also have the highest average cosine similarity among their tweets; a precise inspection of these users suggests that most of them behave as be bots, regularly posting the same message. Later, in the preprocessing step, the presence of bots in the dataset is tackled.
Finally, the *text* field collects the tweet messages. The mes-

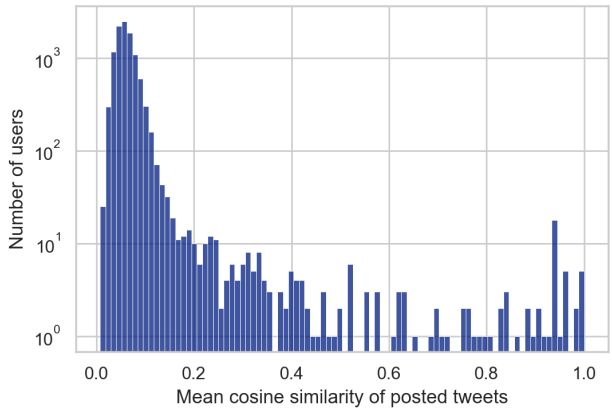

Fig. 3. Histogram representing the number of users per mean cosine similarity computed on the posted tweets of each user

sages are written in english, they may contain tags (i.e., starting with "@"), hashtags (i.e., starting with "#"), and URLs. HTML character entities (e.g., "*&*", "*"*") are present too.

The *text* field is the only one which is going to be considered to train the model and to predict the text sentiment; indeed, the *date* field is only useful to show that both the development set and the evaluation set refer to the same time period; instead, the *user* field is not taken into account because this would instruct the model to predict the sentiment based also on the user who posted the message, loosing the more general scope the model should have.

II. Proposed approach

A. Preprocessing

The model evaluation is made using the evaluation set as reference. As said before, the development set and the evaluation set share the same characteristics, and, presumably, the same source. Basing the model performance on this evaluation set may lead to overfitting, making the model inadequate to predict the tweet sentiment outside this project. Therefore, two possible preprocessing paths have been designed: one more complex, composed of several steps, and more suitable to for a general application; and a much simpler preprocessing path, consisting just in the first two steps of the complex one, that performs better, but probably less general.

The message cleaning steps composing the more articulate preprocessing path are the following

1) "*&*" HTML entities are removed;
2) "*"*" HTML entities are removed;
3) Words staring with @ are removed: these words are user tags, and do not give insights regarding the sentiment of the message;
4) Words staring with *http* are removed: these words are URLs, and do not give insights regarding the sentiment of the message;
5) Punctuation is removed;
6) Stemming is performed, using the Snowball stemmer offered by the nltk module;
7) Negations in sentences are stressed appending in subsequent words a "*_NEG*" suffix.

After these steps, tf-idf is applied on

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. 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).

References

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