UNIVERSITÀ DI PISA

Monitoring the Public Opinion about the Coronavirus Topic from Tweets Analysis

Data Mining project, a.a. 2019/2020

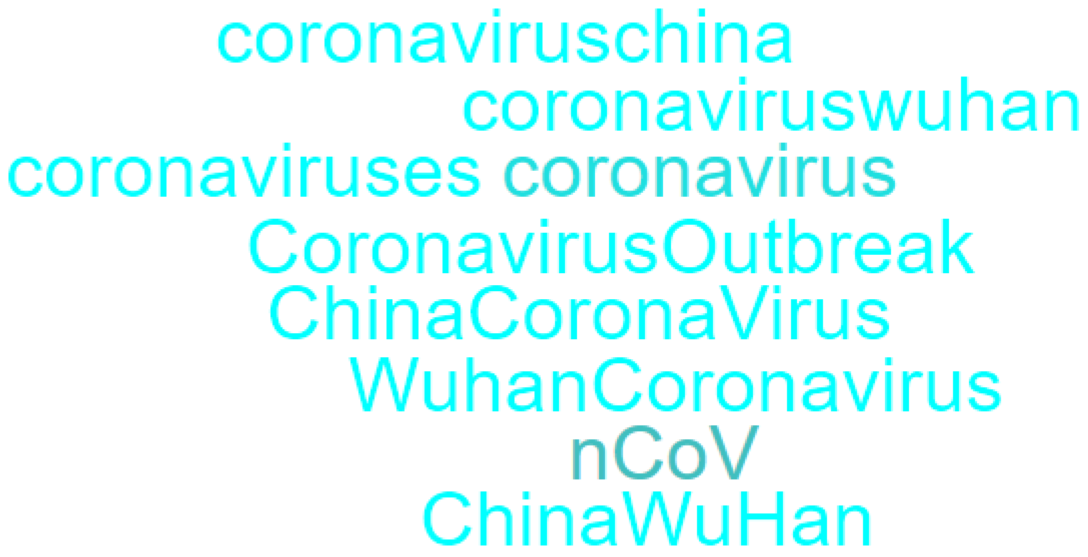
PULIZZI MAURIZIO

WHAT THE APPLICATION DOES

- Measure the level of interest and alarmism regarding the coronavirus topic
- Target: Italian population
- Source of informations: Twitter
- Tweets are classified in three categories:
    - ALARMISTS
    - REASSURING
    - NEUTRAL

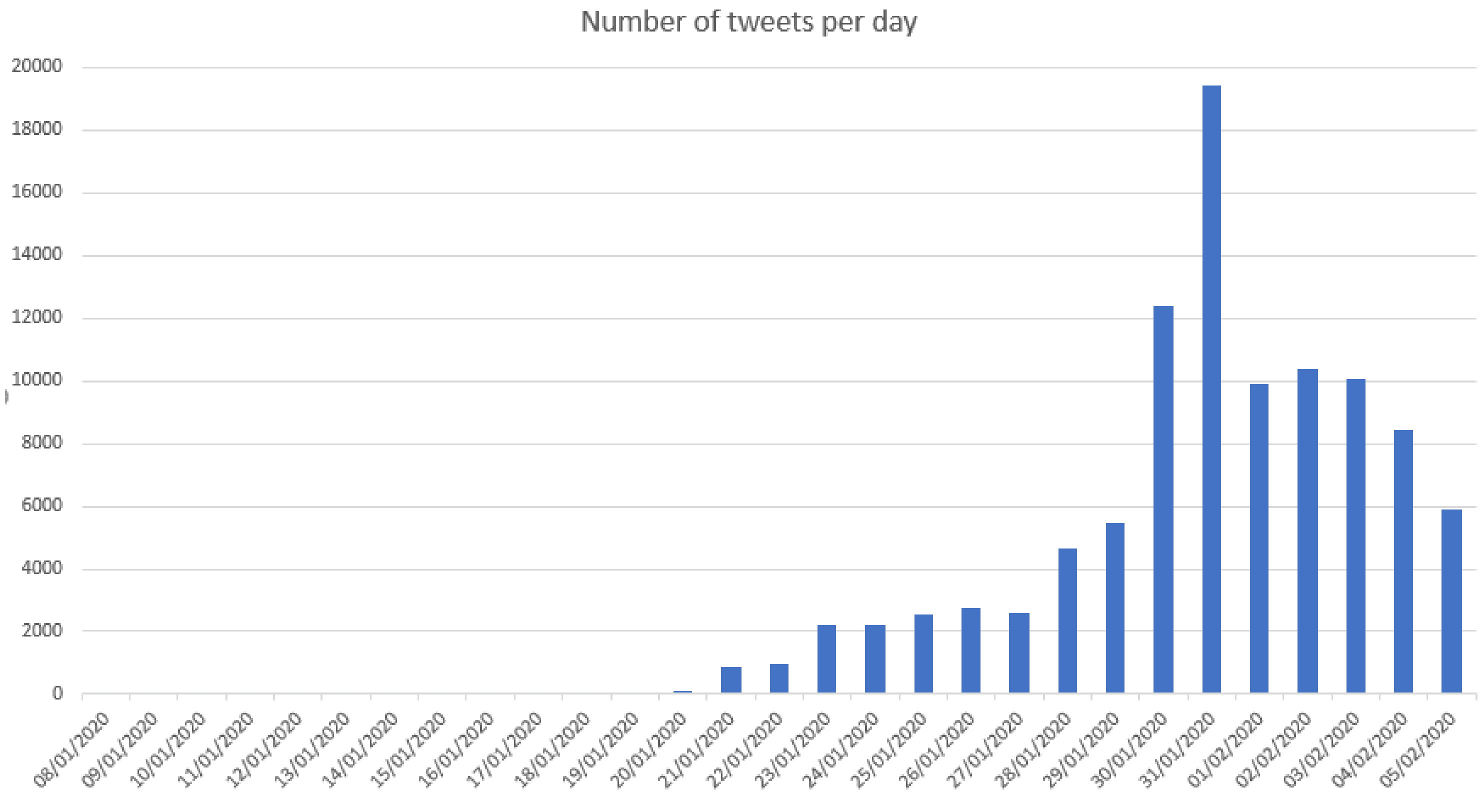
TWEETS OF INTEREST

- Written in Italian
- Having keywords and hashtaas related to the coronavirus



THE TRAININ GSET

Total of tweets captured per days



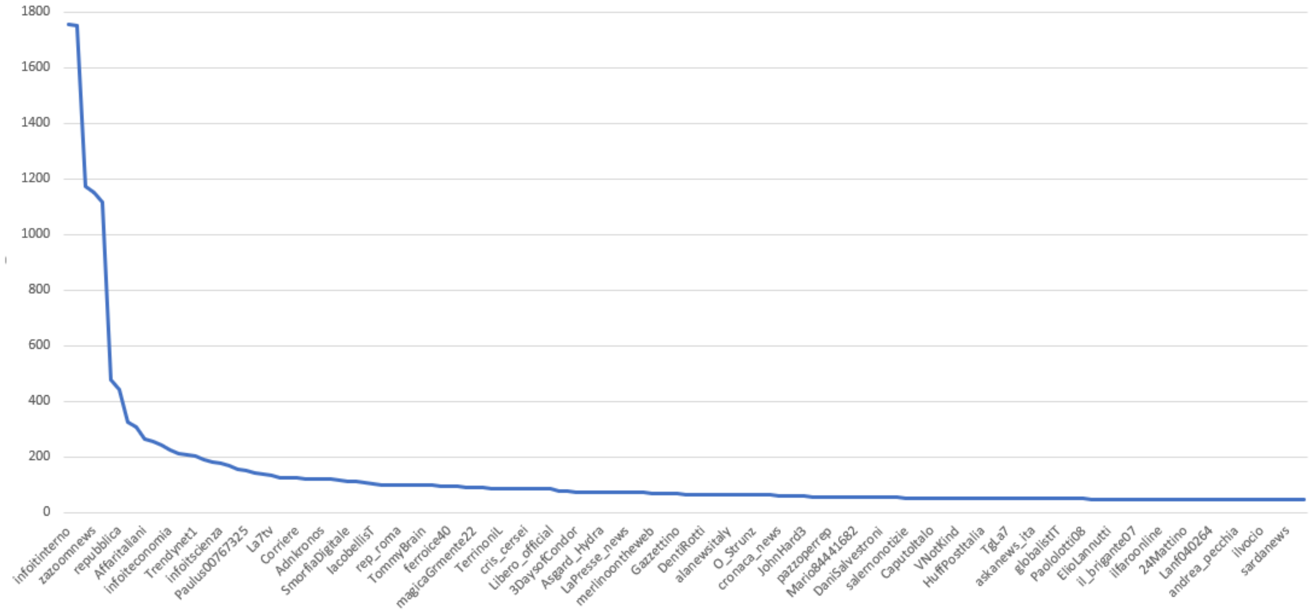
THE TRAININ GSET

○ About 35.000 users in the database

○ 0.6% of most active users generate 20% of the tweets

○ 96% of users wrote less than 10 tweets on the coronavirus in the time frame considered

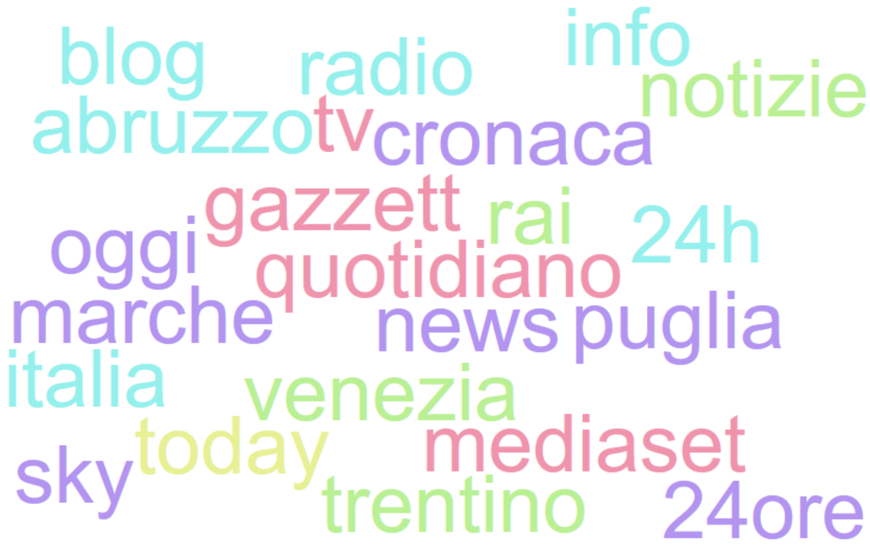
○ 86% of users wrote less than 4 tweets on the coronavirus in the time frame considered



ENHANCEMENT OF THE FILTERING PROCESS ➔ BLACK-LIST OF USERS

From the analysis are excluded all tweets generated by:

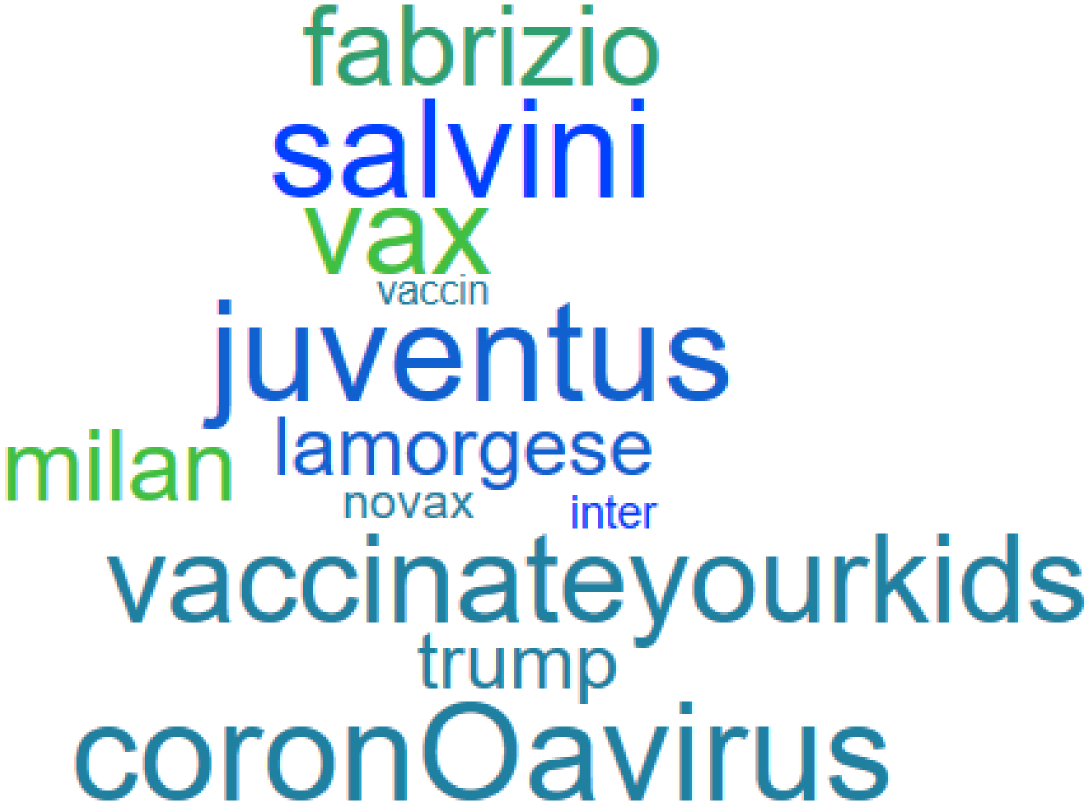
- the 170 most active accounts (accounts that have written more than 43 posts in the range in question)
- 40 other accounts selected from position 171 to 450 in the ranking of most active users
- All accounts that contain words in their username that indicate information pages



ENHANCEMENT OF THE FILTERING PROCESS ➔ BLACK-LIST OF USERS

- Manual labeling was carried out in two phases:
  - 1st phase <u>without</u> user’s black-list
  - 2nd phase <u>with</u> user’s black-list
- more than 80% of tweets viewed in the first phase and belonging to the accounts in the blacklist were neutral.
- in the second phase the percentage of neutral tweets was more than halved

ENHANCEMENT OF THE FILTERING PROCESS ➔ BLACK-LIST OF TERMS IN THE TEXT



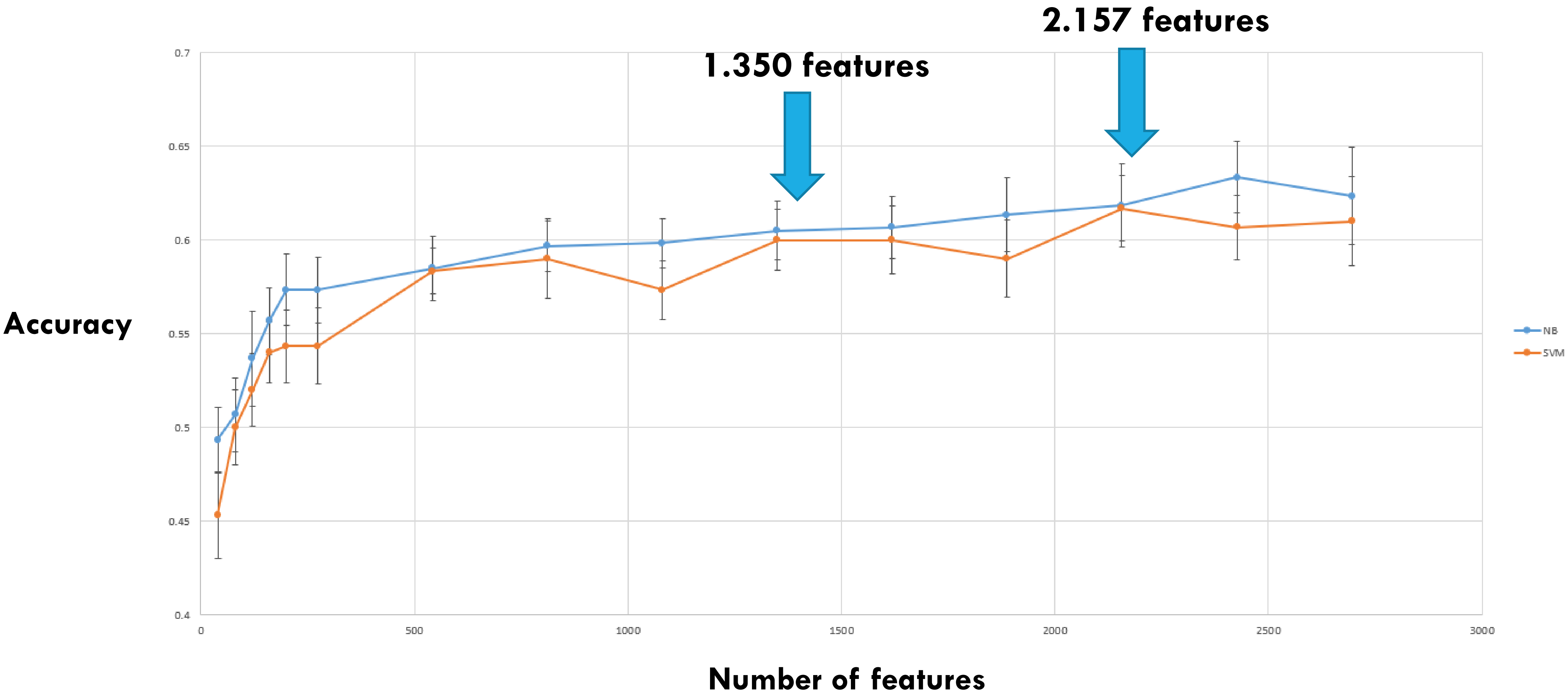
DATA PREPROCESSING

1. Transformation of tweets into lower cases
2. Tweet filtering based on username and words blacklist
3. Tokenization
4. Removal of numbers and hashtags
5. Link truncation, only the initial part is left, i.e. “http”
6. Stop-word filtering (used Snowball-Tartarus list)
7. Stemming (used Snowball-Tartarus stemmer)
8. BOW representation
9. TF-IDF transformation
10. Feature selection (The Information Gain is used to select the features)

CLASSIFIER TRAINING

- Perfectly balanced trainingset (200 tweets per class)
- Tested Naive Bayesian (NB) and Support-Vector Machine (SVM) algorithms
- Tested different number of features (terms)
- Used stratified 10-fold cross validation for compare the results

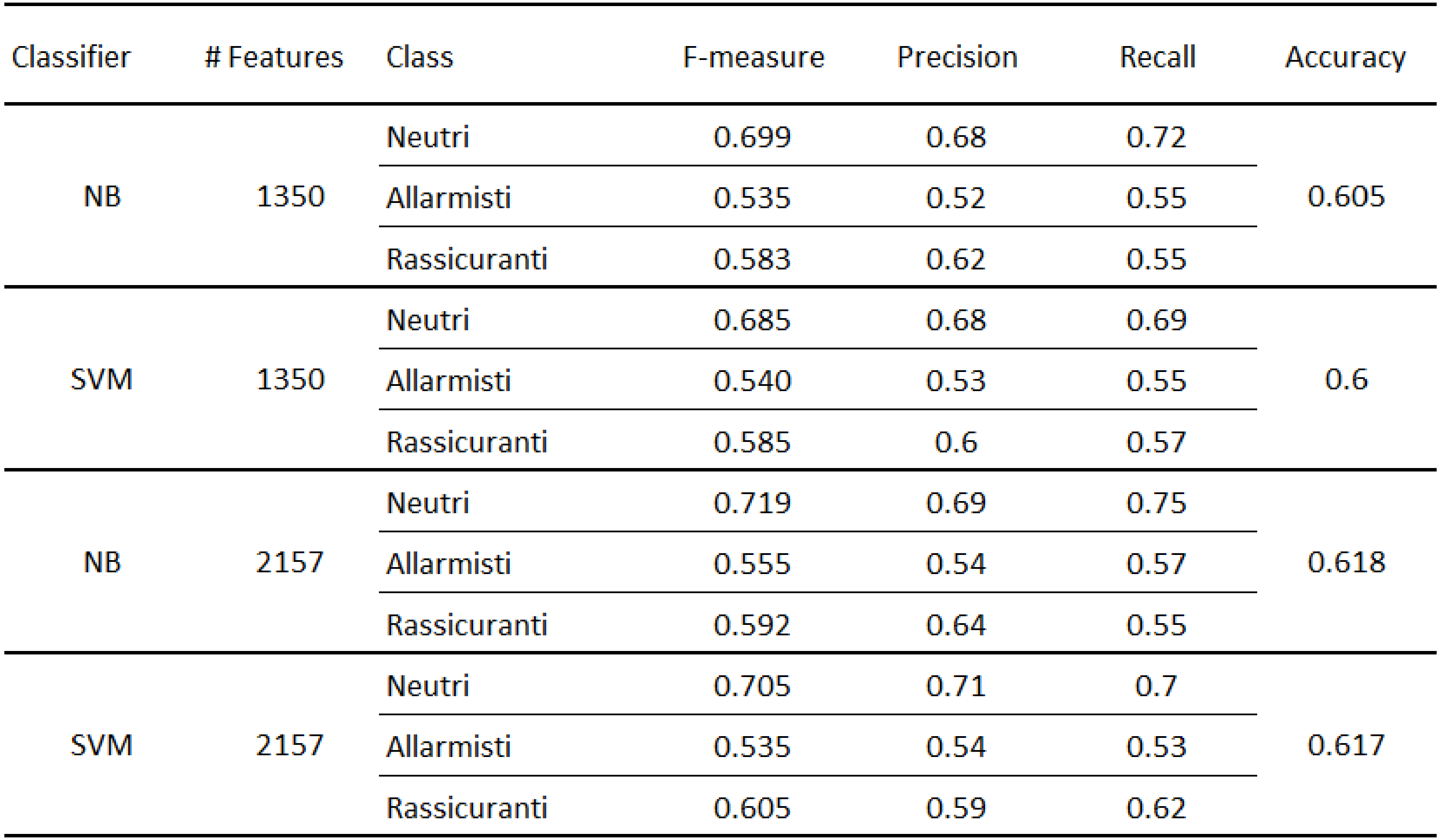
CLASSIFIER TRAINING



2.157 features

1.350 features

TRAINING RESULTS



Classifier	# Features	Class	F-measure	Precision	Recall	Accuracy
NB	1350	Neutri	0.699	0.68	0.72	0.605
		Allarmisti	0.535	0.52	0.55	
		Rassicuranti	0.583	0.62	0.55	
SVM	1350	Neutri	0.685	0.68	0.69	0.6
		Allarmisti	0.540	0.53	0.55	
		Rassicuranti	0.585	0.6	0.57	
NB	2157	Neutri	0.719	0.69	0.75	0.618
		Allarmisti	0.555	0.54	0.57	
		Rassicuranti	0.592	0.64	0.55	
SVM	2157	Neutri	0.705	0.71	0.7	0.617
		Allarmisti	0.535	0.54	0.53	
		Rassicuranti	0.605	0.59	0.62	