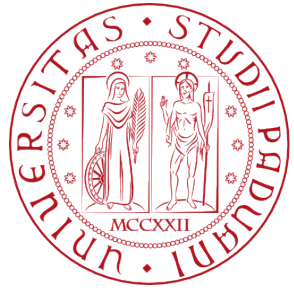
Predicting political orientation from Twitter contents

Alberto Bernardi, Gianmarco Cracco, Elena Izzo,
Luca Lezzi, Alessandro Manente

8 January 2020



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Summary

- Data Collection and Database Creation
- Data Preprocessing
- Model Creation through fastText
- Results
- M5S Prediction

Data collection

- 5 italian political parties
- 1000 italian users randomly selected from each of the 5 parties Twitter's followers
- 3000 Twitter's users randomly selected

Database Creation

- Each user was classified by each member of the groups as a supporter of one of the 5 parties or as Non-Classifiable (NC)
- Data collected by each group were merged together, leading to a $\approx$ 8000 users dataset

Twitter ID	Username	Sex	SID1	SID2	SID3	SID4
Ruggier98623365	Ruggiero	M	1	1	1	1
EricSalicetti	Eric Salicetti	M	0	0	0	0
MassimoDolore	Massimo Dolore	M	4	4	4	4
Rivalevante	Sergio Stagnaro	M	2	0	2	2
Iolanda07658172	Islanda	U	1	3	1	3
manuelamimosa	Manuela Mimosa	F	2	2	4	2

Aim: to construct a model through fastText that is able of predicting whether a user belongs to the right parties (*Lega*, *Fratelli d'Italia*, *Forza Italia*) or to the left ones (*Partito Democratico*)

- Label at 100% vs Label at 75%
- What about *Movimento Cinque Stelle*?

Data Preprocessing

The most important step was cleaning up text data.

Among others, we removed:

- Punctuation
- Stopwords
- Emoji and links
- Insignificant words

We also performed *Lemmatization*.

Model Creation

To create our NLP models we chose fastText.

We trained models in two different ways:

- Labelling each user and considering all their tweets together
- Labelling each tweet of a user with their same orientation

In order to overcome the variability and the unbalance of the dataset, we tried performing the k-fold cross validation, the stratification and the upsampling.

Result - Basic

Parameters			Threshold = 75%	Threshold = 100%
NC	Lemmatization	Person	Accuracy	Accuracy
✗	✗	✗	0,787	0,790
✗	✗	✓	0,699	0,701
✗	✓	✗	0,781	0,786
✗	✓	✓	0,697	0,699
✓	✗	✗	0,653	0,651
✓	✗	✓	0,627	0,599
✓	✓	✗	0,648	0,646
✓	✓	✓	0,617	0,598

- NC: if ✗, then the possible lables are just *DX* (*Right*) and *SX* (*Left*), while if ✓, then there are three possible lables, *DX*, *SX*, and *NC*
- Person: if ✗ the model classifies each tweet, while if ✓ it classifies the single persons considering all their tweets

Result - Upsampled

Parameters			Threshold = 75%	Threshold = 100%
NC	Lemmatization	Person	Accuracy	Accuracy
✗	✗	✗	0,706	0,694
✗	✗	✓	0,623	0,576
✗	✓	✗	0,685	0,679
✗	✓	✓	0,650	0,611
✓	✗	✗	0,605	0,595
✓	✗	✓	0,535	0,543
✓	✓	✗	0,596	0,591
✓	✓	✓	0,530	0,533

- NC: if ✗, then the possible lables are just *DX* (*Right*) and *SX* (*Left*), while if ✓, then there are three possible lables, *DX*, *SX*, and *NC*
- Person: if ✗ the model classifies each tweet, while if ✓ it classifies the single persons considering all their tweets

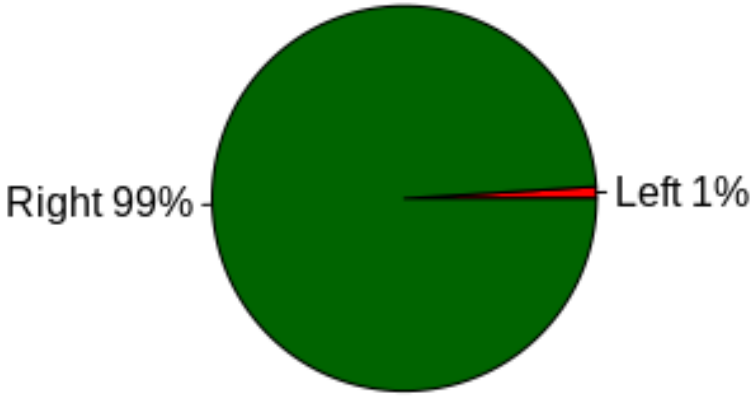
K-Fold Cross Validation

Parameters			Threshold = 75%	Threshold = 100%
NC	**Lemmatization**	**Person**	**Accuracy**	**Accuracy**
✗	✗	✗	0,943	0,801
✗	✗	✓	0,971	0,971
✗	✓	✗	0,899	0,949
✗	✓	✓	0,970	0,970
✓	✗	✗	0,609	0,611
✓	✗	✓	0,648	0,621
✓	✓	✗	0,621	0,623
✓	✓	✓	0,645	0,679

- NC: if ✗, then the possible lables are just *DX* (*Right*) and *SX* (*Left*), while if ✓, then there are three possible lables, *DX*, *SX*, and *NC*
- Person: if ✗ the model classifies each tweet, while if ✓ it classifies the single persons considering all their tweets

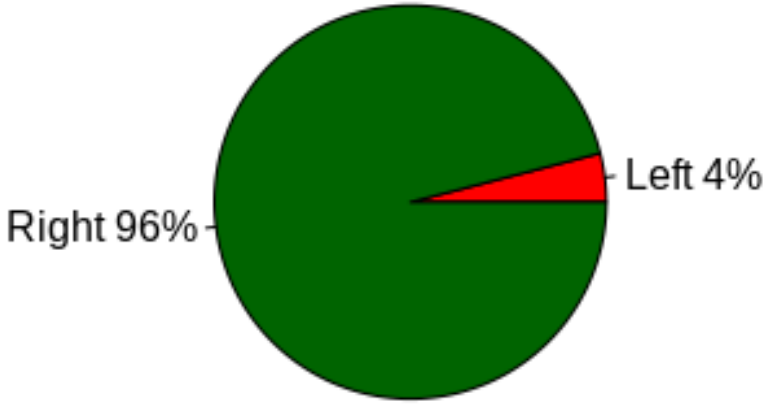
Results - M5S users

How does the best models classify *Movimento Cinque Stelle* users?



Label 75%

- NC ✗
- Lemmatization ✗
- Person ✓



Label 100%

- NC ✗
- Lemmatization ✗
- Person ✓

Thanks for your attention.