Università
degli Studi di
Messina

The textual analysis pipeline for company data: from the sample extraction to the model calibration

Paolo Mustica

Symposium on NLP for company data

17 January 2024 – Utrecht University, Netherlands

Who I am

- **Education**
  - Master's degree in Economics and Finance at the University of Messina, Italy (March 2020). Final grade: 110/110 *cum laude*
  - PhD in Economics, Management and Statistics at the University of Messina, Italy (January 2024). Final grade: Excellent *cum laude*
- **Research interests**
  - Health economics → Alibrandi A., Gitto L., Limosani M. and Mustica P. (2023), Patient satisfaction and quality of hospital care, in Evaluation and Program Planning, vol. 97
  - Regional economics → Limosani M., Millemaci E., Mustica P. (2023), The impact of Special Economic Zones on Southern Italy, Mimeo
  - Textual analysis → Limosani M., Millemaci E., Mustica P. (2023), An efficient Bayes classifier for word classification: an application on the EU Recovery and Resilience Plans, Mimeo

Introduction

- The analysis of unstructured texts has gained importance in Economics in the last few years. Focusing on business data, textual analysis is often applied to investigate the coverage of Environmental, Social and corporate Governance (ESG) topics in business reports
- Textual analysis can provide the useful service of transforming the information of a text into a numerical one, thus making possible statistical analysis. Since we are talking about unstructured data, it becomes important to standardize the procedures for extracting information from the text
- This presentation will focus on the non-financial reports of companies included in the Italian market index (FTSE MIB) to discuss the various steps of the textual analysis pipeline, from the extraction of the sample on which the classification model will be tested to the calibration of the model

Literature review

Summary of the recent literature on the coverage of ESG topics in textual data

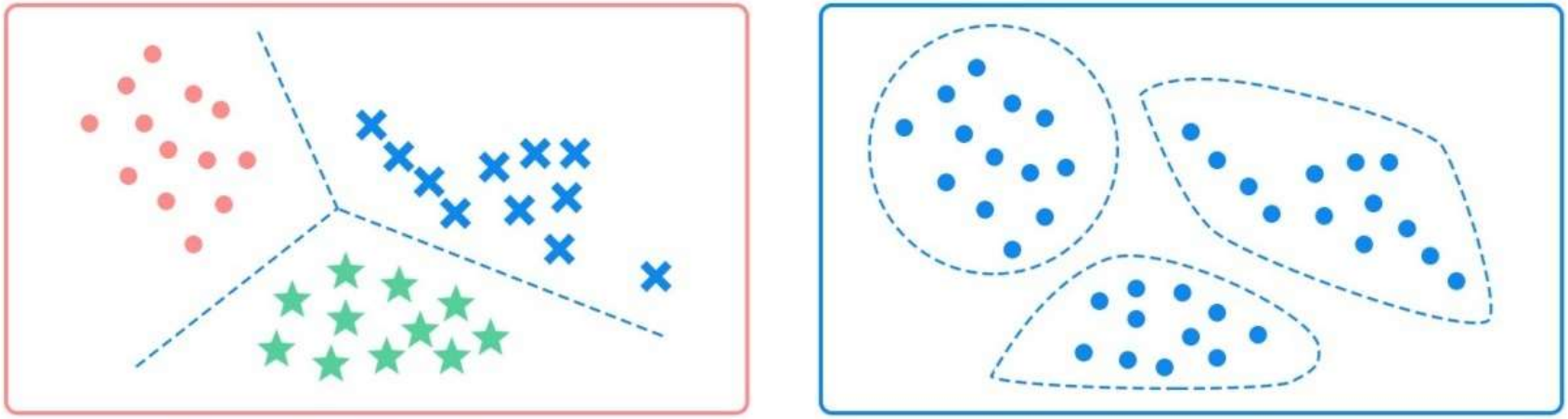
Authors	Text analyzed	Country	Period	Method
Andrikogiannopoulou *et al.* (2022)	Mutual Fund Prospectus	USA	2011-2020	Dictionary based approach
Baier *et al.* (2020)	Form 10-K	USA	2012-2015	Dictionary based approach
Capelle-Blancard and Petit (2019)	ESG news (see the Covalence database)	USA	2002-2010	Labeling by the Covalence team
Heichl and Hirsch (2023)	Non-financial reports according to the NFRD	France, Germany, Italy, Sweden	2017-2021	Dictionary based approach
Kiriu and Nozaki (2020)	CSR reports	Japan	1999-2016	Visualization model based on neural networks
Zeidan (2022)	WhatsApp group of finance professionals	Mostly Brazil	2017-2020	Dictionary based approach

The Italian case study

- As an example, this presentation will focus on the coverage of ESG topics in non-financial reports of Italian businesses
- Non-financial reports are prepared in compliance with the Non-financial Reporting Directive (Directive 2014/95/EU, NFRD). Under the NFRD, large European listed companies are required to publish reports on the policies they implement in relation to social responsibility and environmental matters
- In the case of Italy, we will focus on companies included in the FTSE MIB, the market index that includes the 40 most traded companies on the Italian stock exchange. In particular, we will investigate the non-financial reports of these companies published between 2017, when the publication of such reports became mandatory, and 2022

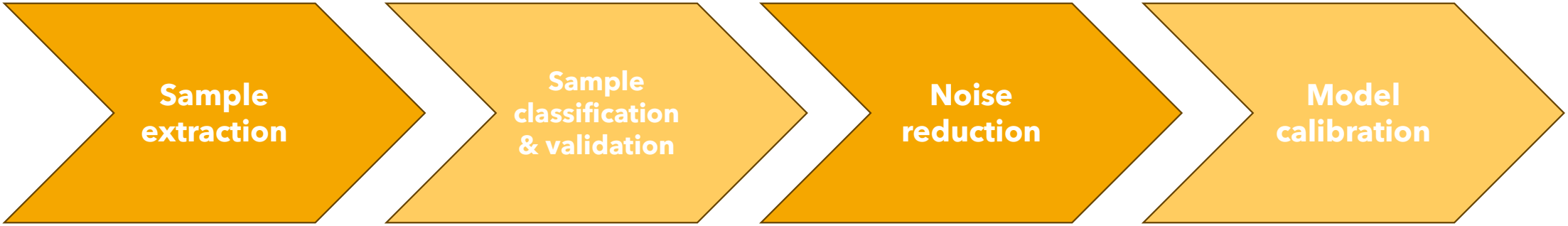
Machine Learning approaches

- The textual analysis pipeline discussed in this presentation refers specifically to supervised methods



Pic adapted from: Supervised vs. Unsupervised Learning: What's the Difference?, Smriti Saini, 2021

Textual analysis pipeline



Sample extraction

- The documents to be analyzed are usually unstructured data. Consequently, textual units contained in these documents (n-grams, sentences, ...) are not already labeled in classes. This means that we cannot calibrate the model we want to use on the classes to be investigated
- To handle this problem, we can manually label a sample of textual units. In particular, we need a random sample to be sure that what we are observing in the sample represents the whole population of textual units
- Since we can represent the population as an ordered list of N textual units, we can use systematic sampling to generate a random sample of n textual units using the sampling interval k

$$k = \frac{N}{n}$$

Classification and validation of the sample

- The next step is labeling each unit in the random sample in the corresponding class. It is one of the most delicate parts, because the labeling is influenced by the subjectivity of the researcher. This issue can be mitigated in two ways
  1. the content of the classes must be clear to the researcher (it is important to understand what the literature says about them)
  2. independent validators can improve the quality of the original labeling: in case of discrepancy between the labeling of the researcher and the labeling of the validator, we have to decide on a case-by-case basis whether to maintain the original labeling or whether to change it with the labeling of the validator
- We have to split the random sample into training set, which is used for training the classification model, and test set, which is used for testing the classification performance of the trained model. The training set is typically between 70% and 80% of the random sample

The validated random sample

Number of training and test sentences (clusters of words) labeled as E, S and G

Topic/Set	Training	Test	Total
E	371	150	521
S	474	138	612
G	1452	421	1873
Total	2297	709	3006

Noise reduction

- Before calibrating the classification model, we have to reduce the noise in textual units. Indeed, natural language is characterized by a lot of words that can confuse the classifier. There are several strategies to handle this problem
  1. Train your classifier on the vocabulary actually used in the documents to be analyzed: although the English vocabulary includes about 170 thousand words, technical documents usually use a few thousand words
  2. Remove numbers and stop words: numbers and stop words (commonly used words from which we cannot extrapolate useful information) are useless for classification purposes. For this reason, we can remove them
  3. Reduce words to their word stem. Known as stemming, this process reduces multiple words to their word stem, allowing us to shrink the vocabulary (as an example, the words “change”, “changed” and “changing” can be reduced to the word stem “chang”)

Calibration of the classification model

- Now we can start with the calibration of the model. In general, the calibration is characterized by the following steps
  1. Train the model. The training set is used to train the model and test it on the test set
  2. Choose the model parameters. Usually each model is characterized by one or more parameters. We have to set them in such a way that we get good classification performance on the test set
  3. Augment data and/or regularize parameters. Data augmentation and regularization are strategies usually used by machine learning practitioners to increase the classification performance, especially in those classes where we obtain poor results. While data augmentation is a technique of artificially increasing the training set, regularization shrinks the weight of parameters to avoid overfitting
- Let's look at these steps using the Prior Adaptive Bayes (PAB) classifier (Limosani *et al.*, 2023)

The PAB classifier in a nutshell

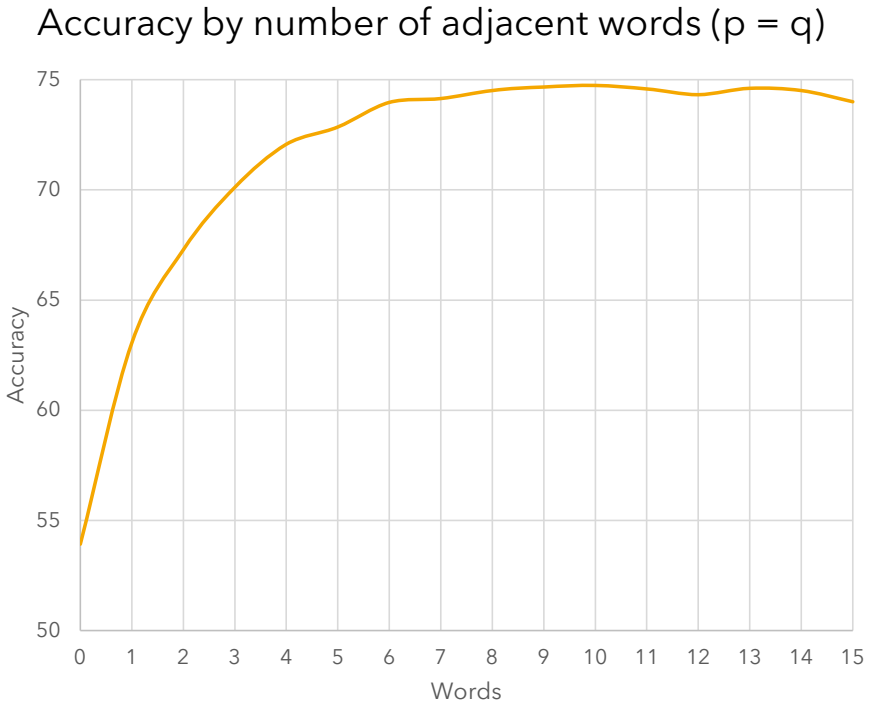
- The PAB classifier is an adaptation of the Bayes classifier for the word classification task, which exploits the rule that words in a text are clustered in topics (topic clustering assumption). In particular, it allows us to achieve good classification results at the word-level without being computationally expensive
- Its main characteristic is that the prior probabilities of classes, $pr(c_j)$, are replaced with the corresponding posterior probabilities given that the previous p and the next q words to the i-th word to be classified occurred, $pr(c_j|x_{i-p} \cap \cdots \cap x_{i+q})$

Bayes classifier	$pr(x_i = w_k\|c_j)pr(c_j)$
PAB classifier	$pr(x_i = w_k\|c_j)pr(c_j\|x_{i-p} \cap \cdots \cap x_{i-1} \cap x_{i+1} \cap \cdots \cap x_{i+q})$

- where x_i is the i-th word in the text, w_k is the k-th word in the vocabulary and c_j is the j-th class. Each x_i will be classified in the class with the maximum product between the two probabilities

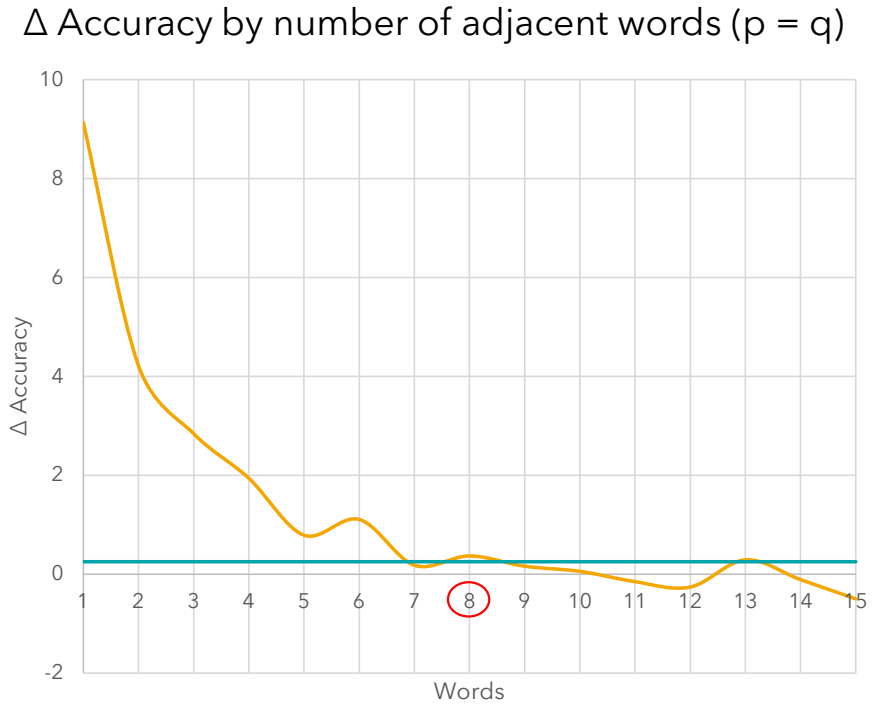
PAB classifier – Training

- The training of the PAB classifier is based on the computation of the likelihoods, $pr(x_i = w_k | c_j)$
- In particular, a likelihood is given by the relative frequency of word k in the vocabulary labeled with class j in the training set
- Once trained, the classifier can be tested: its accuracy increases as p and q increase, but at decreasing rates



PAB classifier – Choice of parameters

- It is better to choose the number of adjacent words sparingly to avoid overfitting on the sample
- We can choose this number by considering the improvement in terms of classification compared to the previous number of adjacent words
- When this improvement becomes negligible, we can stop. After p = q = 8 it is systematically less than 0.25%



PAB classifier – Data augmentation (1/2)

Local accuracies

- Although with p = q = 8 we obtain a good accuracy, we should look at local accuracies to understand if we need to augment the training set in a class

	Total	E	S	G
	74.51	69.94	67.54	79.14

Augmentation routine

- We augmented n times (epochs) the training set in the classes E and S to get a more balanced accuracy among classes

FOR x in 1 to n
Find the top 10 misclassified E and S words in the epoch x-1
Increase the number of such words in the training set of epoch x-1, getting the training set of epoch x
Train the PAB classifier with the training set of epoch x
Run the PAB classifier for p = q = 8
END FOR

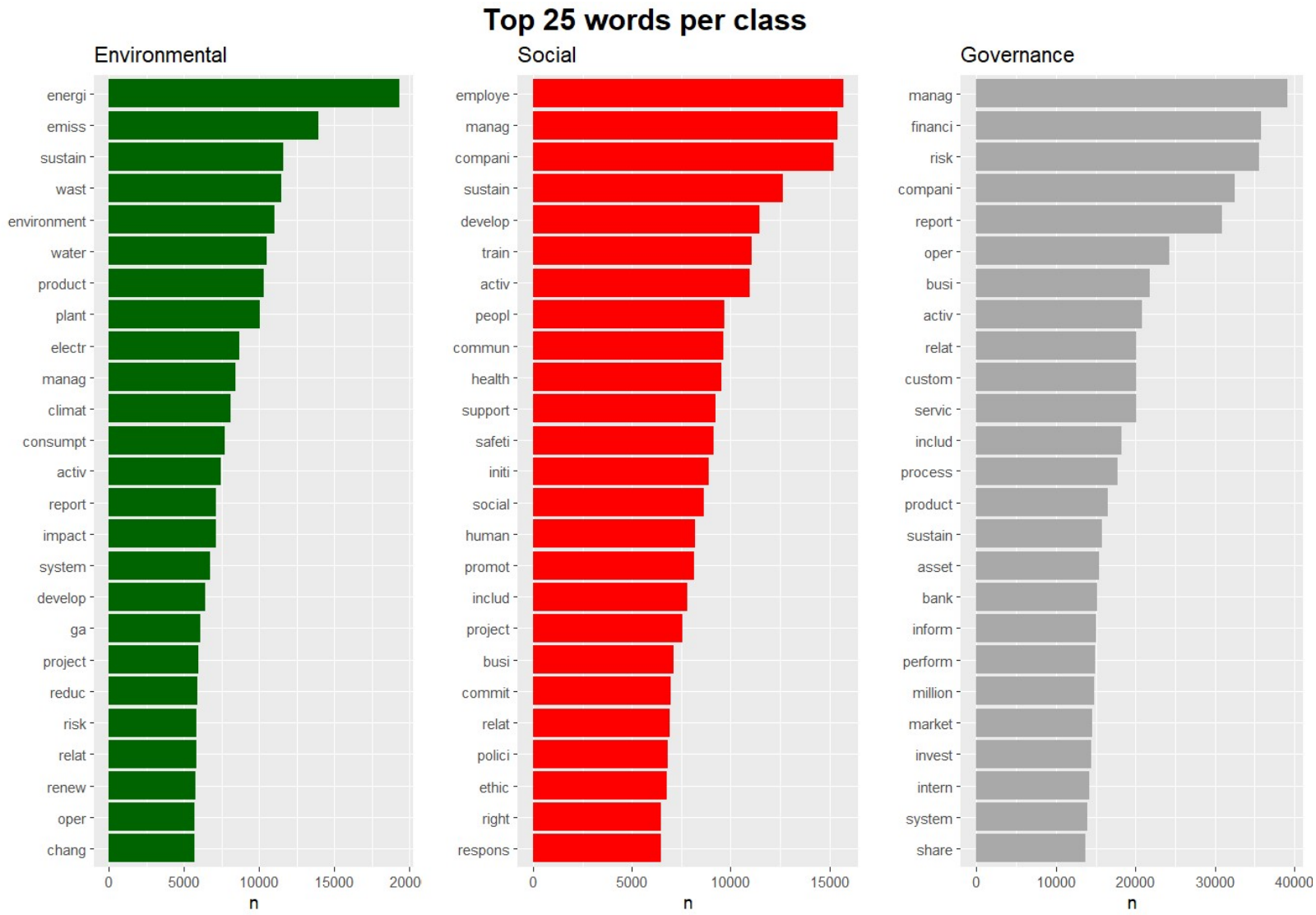
PAB classifier – Data augmentation (2/2)

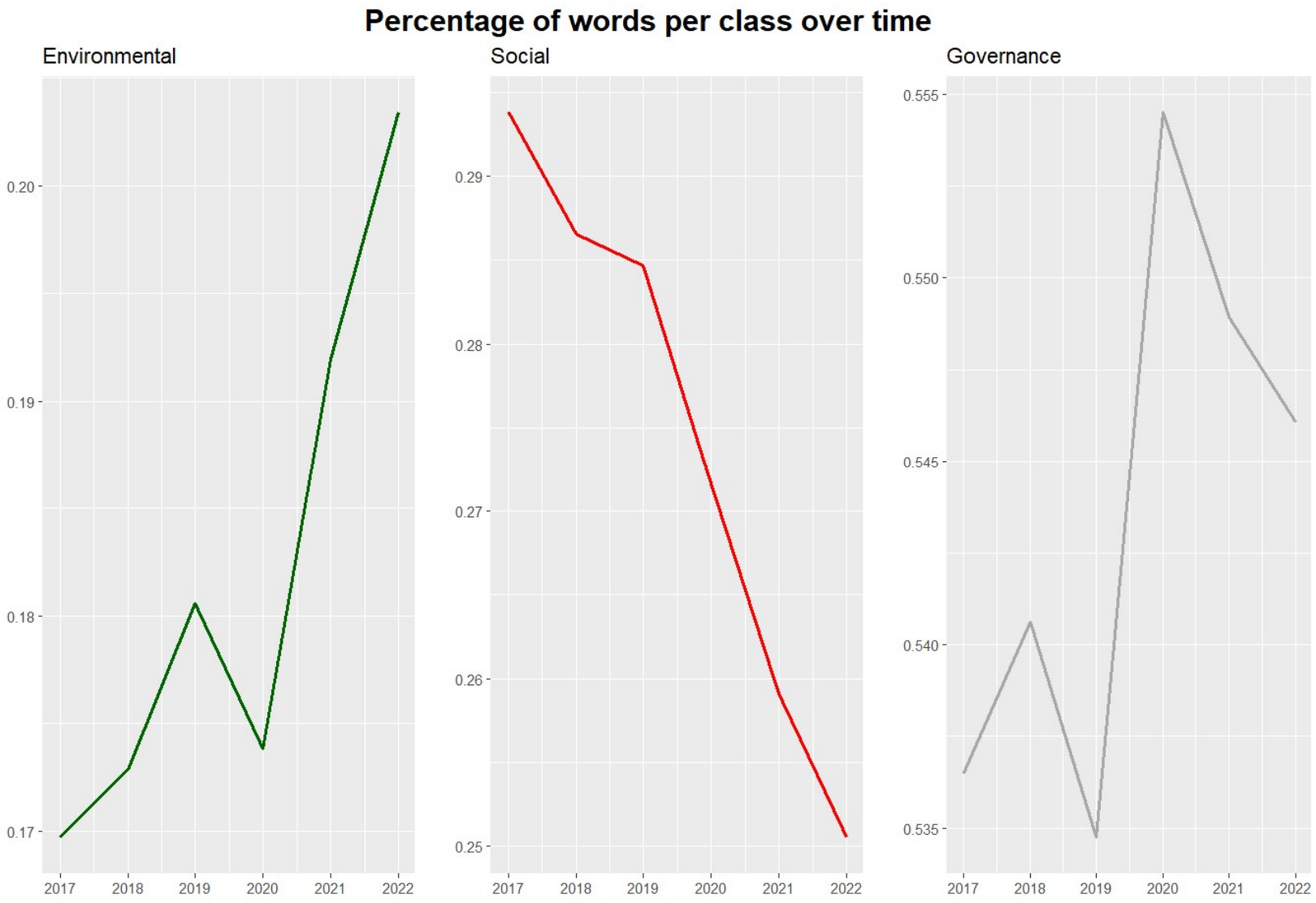
- At the epoch 3 we obtained the lowest average deviation of E and S compared to G

Local accuracies (p = q = 8) by epoch

Epoch/Accuracy	Total	E	S	G	Average deviation compared to G
0	74.51	69.94	67.54	79.14	10.40
1	75.07	71.75	69.69	78.55	7.84
2	75.22	72.44	70.99	78.04	6.32
3	75.68	73.29	71.48	78.30	5.92
4	75.66	73.56	70.74	78.44	6.28
5	76.03	74.31	70.14	78.99	6.76

- Once calibrated the model, we can run it on the whole population of textual units


Top 25 words per class
Environmental
energi
emiss
sustain
wast
environment
water
product
plant
electr
manag
climat
consumpt
activ
report
impact
system
develop
ga
project
reduc
risk
relat
renew
oper
chang
0
5000
10000
15000
2000
n
Social
employe
manag
compani
sustain
develop
train
activ
peopl
commun
health
support
safeti
initi
social
human
promot
includ
project
busi
commit
relat
polici
ethic
right
respons
0
5000
10000
15000
n
Governance
manag
financi
risk
compani
report
oper
busi
activ
relat
custom
servic
includ
process
product
sustain
asset
bank
inform
perform
million
market
invest
intern
system
share
0
10000
20000
30000
40000
n


Percentage of words per class over time
Environmental
Social
Governance
0.20
0.19
0.18
0.17
0.29
0.28
0.27
0.26
0.25
0.555
0.550
0.545
0.540
0.535
2017
2018
2019
2020
2021
2022

Possible hypotheses to test

- Textual information can be used for testing several hypotheses. As an example:
  - Do businesses that perform worse in environmental indicators talk more about environmental issues (green washing hypothesis)?
  - Are there differences in the coverage of ESG topics based by sector?
  - Are businesses that are more involved in environmental and social matters rewarded by investors?

References

- Andrikogiannopoulou A., Krueger P., Mitali S. F. and F Papakonstantinou (2022), Discretionary Information in ESG Investing: A Text Analysis of Mutual Fund Prospectuses, SSRN
- Baier P., Berninger M. and Kiesel F. (2020), Environmental, social and governance reporting in annual reports: A textual analysis, Financial Markets, Institutions & Instruments, vol. 29, issue 3, pp. 93-118. doi: https://doi.org/10.1111/fmii.12132
- Capelle-Blancard G. and Petit A. (2019), Every little helps? ESG news and stock market reaction, Journal of Business Ethics, vol. 157, pp. 543-565. doi: https://doi.org/10.1007/s10551-017-3667-3
- Heichl V. and Hirsch S. (2023), Sustainable fingerprint – Using textual analysis to detect how listed EU firms report about ESG topics, Journal of Cleaner Production, vol. 426, 138960. doi: https://doi.org/10.1016/j.jclepro.2023.138960
- Kiriu T. and Nozaki M. (2020), A Text Mining Model to Evaluate Firms' ESG Activities: An Application for Japanese Firms, Asia-Pacific Financial Markets, vol. 27, pp. 621-632. doi: https://doi.org/10.1007/s10690-020-09309-1
- Limosani M., Millemaci E. and Mustica P. (2023), An efficient Bayes classifier for word classification: an application on the EU Recovery and Resilience Plans, Mimeo
- Zeidan R. (2022), Why don't asset managers accelerate ESG investing? A sentiment analysis based on 13,000 messages from finance professionals, Business Strategy and the Environment, vol. 31, issue 7, pp. 3028-3039. doi: https://doi.org/10.1002/bse.3062

Thank you for your attention!

pamustica@unime.it