Università
Ca'Foscari
Venezia

Master's Degree

in Economics and Finance

Final Thesis

ESG valuation: a web scraping approach

Supervisor
Ch. Prof. Loriana Pelizzon

Assistant supervisor
Ch. Prof. Shouro Dasgupta

Graduand
Simone Obrizzo
Matriculation number
877797

Academic Year
2019 / 2020

Contents

Preface

Since the ESG (environmental, social and governance) topic is gaining more attention with Governments encouraging “corporate stewardship” [1], companies have decided to provide useful information and proactively approach the ESG topic. According to Deegan [2] [3] the reasons companies redact ESG reports are regulation and attention to stakeholders. These reports are also used to match the requirements some investors and customers have. By aligning the objective, the companies will gain support and legitimacy that is a strategic resource for future growth [4].

The 2020 Trends Report from the Forum for Sustainable and Responsible Investment [5] shows how the assets under management involved in ESG investment strategies grew 42% in the period 2018-2020 and 38% from 2016 to 2018.

Along with the number of investments, it is also increased the number of ESG reports that have been produced by Sustainability Rating Agencies (SRAs) [6]. The rating methods proposed, and the calculation processes are not univocal, therefore, they are not leading to similar scores. In particular the measurement disagreements are pronounced on Human Rights and Product Safety [7]. In addition to that, different ratings have consequently scattered the decision of investors that, given the difference in preferences, do not have a statistically significant impact on financial performances [8].

In this framework I have decided to analyze companies approach to the sustainability theme through a web crawling, web scraping method. This thesis investigates the relation between how a sample of Italian companies market themselves on their websites and their actual ESG valuations. Moreover, it is analyzed if there is a relation between ESG web scraped scores and financial measures.

The study is performed through the analysis of the companies’ commitment to explain to their clients and shareholders the steps taken to face the environmental, governance and social (ESG) challenges. A web scraping algorithm that targets the websites of the Italian companies is used to determine the level of attention to the topics cited above and assigning them new scores. The crawler and the scraper [9] are written in Python programming language with the Scrapy framework. They perform the task of collecting URLs and all the text coming from the targeted company’s web pages that have in the uniform resource locator (URL) the word “sostenibilita” or similar. Given the different structure of every site and the different programming languages with which they are written, if the previous process is not possible or the data collected are insufficient for a comparative analysis a new code is written for the specific site.

I then created my own dictionary by using lexicons from different studies [10] [11] and confronting them with the most used Italian words in the ESG space. To each word in the dictionary is assigned a score that is used to create a single, new, E, S, G score of the overall website. The ratings will then be confronted with the ratings performed by Bloomberg and then confronted with the data reported in financial reports by each company.

My contribution to the literature with this work is the introduction of a web crawling and scraping algorithm to evaluate companies under the ESG lens. This method has not been used to analyze listed companies' websites with an ESG objective. Also, the innovation stands on the analysis of the relation between the web scraped ESG scores and the reported financial data. Moreover, with the introduction of a new dictionary I created a base for the analysis of the extracted text files in the Italian language.

The thesis starts in Chapter 1 with the analysis of the ESG framework comprehensive of the literature review, in particular is described the relation between the ESG ratings and the financial performances, the cost of capital for different level of sustainability, the European and U.S. legal framework and is presented the S-CAPM. In the following Chapter is discussed the web scraping process with particular attention to its theoretical foundation and the potential on the future of investing. Then are illustrated the characteristics of the most common libraries to scrape the websites' data. Chapter 3 will illustrate the processes conducted in my research. The Python code and the dictionary are presented to evaluate the Italian companies under their ESG profile; results are then confronted with the SRA ratings. Lastly, the relation between the results and the specific financial ratios and performances are evaluated. The last chapter, Chapter 4, draws a conclusion and suggests possible following topics to expand the research question.

Chapter 1
ESG

1.1 Introduction

Socially Responsible Investing (SRI) involves the consideration of ESG (environmental, social and governance) factors for the selection of the securities. While ESG shapes valuation, SRI screens investments positively or negatively, with the ESG factors help [12]. The idea of socially aware investors that use the investment criteria in connection to the personal values and beliefs was a concept that has been initially taken into account by religious groups in 1928 [13]. In the 1970s and even more in the 1980s the movement shifted from a religious clientele to a broader range of investors where the first SRI movements, and specifically anti-apartheid, has been found [14] [15].

In the recent years the number of investors that take into account ESG topics in their investments' decision is increasing and more than two third of all the sustainable investing assets are held by institutional investors [16].

There are two main reasons why a ESG investment strategy could be taken into account by individual and institutional investors: the first one is the higher returns derived from the strategy and the second one is a parallel goal alongside with returns. For the first there is evidence of both outperformance and underperformance based on the chosen criteria, for the second scope is a question of preference that can lead to tension between social, economic justice, environmental protection and profit maximization [17].

In this chapter, I will explain the definition, the state of art of the ESG investing. Then I will provide a deeper look into the causes of different financial performances of corporations and the legislation state in EU and U.S.. In conclusion will be described and analyzed the sustainable CAPM model.

1.2 Environmental pillar

The analysis of the environmental aspect of a company is related to its ability to create production processes that are sustainable for the environment, in particular the attention from the raters is focused on the pollution management, the waste disposal, carbon emissions and the renewable energy usage. Companies that have been reluctant to implement environmental management systems (EMS), or ISO14001, are due to a perception of cost disadvantage, but evidence shows that financial performances are not negatively impacted by the introduction of EMSs [18]. The implementation success is related to the involvement of employees and management, government initiatives and adequate organizational resources. After the adoption of ISO 14001 there are improvements in organizational reputation, in its relationship with their stakeholders, in the organizational effectiveness and in the environmental risks' perception [19] [20].

With the increasing external pressure to environmental corporate social responsibility (CSR) the positive market reaction to eco-friendly actions has been decreasing, while the negative reaction to eco-harmful acts has been increasing. Moreover the market reaction to green initiatives from companies that have an already high level of CSR has been more tempered [21].

1.3 Social pillar

This pillar is related to the social responsibility of a company, and it is evaluated by including all the relation a company has with its customers, employees, suppliers, governments and society. The most adopted policies by corporations include human rights, communities' support, business ethics, data and privacy protection.

Evidence shows that gender-diverse board contributes to reduce firms' risk by reducing firms' stock return volatility [22]. The woman participation on boards was 7.7 percent in 2008, 10.7 in 2013, and 30 percent in 2019 for the companies in the FTSE350. A study from Jizi et al. [23] shows how a group of female directors on the board has a higher negative coefficient to volatility than the coefficient of a dummy variable related to the appoint of a female director.

In 2008 Scholtens & Zhou [24] found not statistically significant relation between social responsible activities of 289 companies and their financial performances. More recent studies on the other hand highlighted how employees' level of satisfaction of the work conditions [25], and customer satisfaction [26] were related to higher corporate financial performances (CFP).

1.4 Governance pillar

The governance pillar is related to the level of transparency in the creation of company's policies, the interaction with shareholders, the correctness of managers' remuneration and the internal control processes.

Nollet et al. [27] found a non-linear relation between corporate social performance (CSP) and stock returns, in particular the relation is found to be U-shaped. While initial CSP investments will lead to lower financial performances, a confirmation is also found in Lopez et al. [28], after a certain amount is spent the investments will pay off. In particular the governance sub-component of the model is found to be the most influential to the improvements of the CFPs.

1.5 Literature review

In April 1972 in an article written by J. Bragdon and T. Marlin [29] the relation between environmental virtue and financial performances started to be addressed and the results were that the two were compatible. They highlighted how a good profit record and pollution controls were also driven by good management practices. In the same year Moskowitz [30] after conducting a research on 14 socially responsible companies and confronting their performances with the Standard & Poor 500 and the Dow Jones found that the first ones outperformed the indexes. Three years later Vance [31] found a negative relation and offered what appears to be the first, but not the last, contradictory view on the subject.

The number of studies since the 1970s has been increasing and accelerating since the 1990s as showed in 2015 by Friede et al. [32] that found that more than 2200 academic studies were written on the relation between ESG and corporate financial performances. In particular 90 percent of studies found a non-negative relation between environment, governance, social responsibilities and financial performances. In another academic literature review conducted in the same year by Clark et al. [33], they found that in 45 studies out of 51 (88 percent) showed positive correlation between returns and sustainability. While for the financial performances the positive correlation was found in 80 percent of the total, 41, cases.

But the results are far from conclusive since evidence shows that ESG-CFP overtime presents a non-clear picture [34] [35] [36] and studies from developing countries are insufficient to draw conclusions [37].

Companies that use CSR practices are better positioned to perform better than the ones that are not focused on these behaviors [38] and during the financial crisis of 2008 this is also shown by Ortas et al. [39] using a Multivariate Generalized Autoregressive

Conditional Heteroskedasticity (M-GARCH) analysis to the stocks included in the FTSE4Good-Ibex[1].

Zhao et al. [40] analyze how China's energy power companies improve their financial performance (ROCE[2]) through the pursuit of good ESG performances. Positive results are also found from Dalal and Thaker [41] after analyzing with the random effect panel data regression analysis the influence of ESG factors on the financial performance of 65 Indian firms listed on the Nifty100 ESG Index[3]. Sahut and Pasquini-Descomps [42] found that the U.S., U.K. and Switzerland stocks return in the period 2007-2011 and the variation of the ESG scores, coming from 8093 monthly observation of change, are not significant except for the U.K.; they also found from their non-parametric kernel regression[4] that the function between the two is not linear. The study from Evans and Peiris [43] confirmed the relation in the U.S. stock market by analyzing the MSCI KLD 400 Social Index[5]. Fischer and Sawczyn [44] investigate the ESG-CFP relation in the German market and found a Granger-casual[6] relation, suggesting a simultaneous, lagged, causality between the two. Velte [45] confirmed the evidence in the German market using data from 2010 to 2014 and by measuring FP with ROA and Tobin's Q.

He found that all the subcomponent of ESG were statistically significant on the return on asset but were not in the Tobin's Q. In Italy Landi and Sciarelli [46] used panel data analysis through a Fixed Effect Model to evaluate if the ESG rating impacted a firm's abnormal returns. In the time period evaluated, 2007-2015, they found a non-statistically significant relation between market premium and the ratings assigned.

[1] The index comprises companies in the BME's IBEX 35 Index and the FTSE Spain All Cap Index that demonstrate good sustainability practices. The criteria to be selected to enter the FTSE4Good are strong ESG risk management practices. In order to be included in the FTSE4Good Index Series companies must have an overall ESG Rating of 3.1 out of 5.

[2] ROCE is the ratio of earnings before interest and tax to the total capital employed by the company.

[3] The Nifty100 ESG Index reflects the performance of companies within NIFTY 100 index, based on Environmental, Social and Governance (ESG) scores. The weight of each constituent in the index is tilted based on ESG score assigned to the company, i.e. the constituent weight is derived from its free float market capitalization and ESG score.

[4] The kernel regression is a non-parametrical method used to estimate the conditional expectation of a random variable that can be written as:

$$E(Y|X) = u(X) \qquad u\ is\ an\ unknown\ function.$$

[5] The MSCI KLD 400 Social Index is a capitalization weighted index of four hundred U.S. securities that provides exposure to companies with outstanding ESG ratings.

[6] The Granger causality test determines if lagged values of time series could be used to predict a second, non-lagged, time series.

There is evidence also of cost of capital differences between the ESG stocks and sin stock, that are the companies involved in what are socially considered immoral businesses, an example are alcohol, gambling and tobacco companies. Hong and Kacperczyk [47] find that sin stocks have higher cost of capital and institutions, especially pension funds, pay a financial cost for not being able to invest in those stocks given also the higher expected returns compared to other types of investable stocks. This difference in expectation is due to possible legal actions these stocks generally face.

1.5.1 Cost of debt in eco-friendly firms

The cost of debt influences a company's level of risk and its profitability. Negative ESG externalities increase reputation and financial risk. Academics' hypothesis is that a higher ESG management reduce the cost of distress by lowering the probability of incurring in end tail events and this would lead to lower cost of debt. In 2008, Sharfman and Fernando [48] analyzed 267 U.S. firms and found a statistically significant positive relation between cost of debt and ESG management practices. Two years after similar conclusions are found by Bauer and Hann [49] by analyzing 2200 corporate bonds they found that better ESG standards were related to lower credit spreads. Chava [50] in the analysis of 1341 firms showed how higher interest rates are paid by companies that have some form of CSR concern. On average the premium paid by these companies is in the range of 7-18 percent more than firms with ESG plans [51].

The amount spent on socially responsible activities has a non-linear, U-shaped relation with the cost of debt; in particular Ye and Zhang [52] found that in China when a company has a very low or very high charitable donation/sales ratio the cost of debt results higher.

1.5.2 Cost of equity in eco-friendly firms

While the result from Sharfman and Fernando [48] showed previously demonstrated a positive relation with the cost of debt, with the cost of equity, calculated by using the CAPM, and with the weighted average cost of capital the relation was negative and statistically significant. But there is also in this case contrasting literature, and by the number of it, is in favor of a positive relations. Dhaliwal et al. [53] found a 1.8 percent discount in the cost of equity for the firms that had optimal ESG measures and decided to disclose them; also, CSR disclosing reduce the forecasting error and the effect of opaque financial reporting [54]. El Ghoul et al. [55] after analyzing a sample of U.S. firms conclude that there is a negative, statistically significant correlation between employees' satisfaction, environmental policies and cost of equity. Albuquerque et al. [56] empirically evaluated that the correlation between a firm's beta and their CSR index is negative and statistically significant.

Moreover, there have been conducted analysis on non ESG-friendly stocks that show that investors require higher returns for holding companies that are not socially responsible. Evidence is found by Girerd-Potin et al. [57] for French investors and for Chinese investors by Li et al. [58] that fund higher cost of equity for bad carbon emission reports. In conclusion the cost of capital has a negative relation with positive CSR reports. The MSCI World Index has calculated that from December 31, 2015, through November 29, 2019 the average cost of capital[7] of the best ESG reporting firms (highest quintile) was 6.16 percent, while for the worst ESG reporting firms (lowest quintile) was 6.55 percent. The spread for the MSCI Emerging Markets is even higher.

Good ESG practices reduce the systematic risk, that is the beta, and consistently with the CAPM they reduce the cost of equity for a firm.

1.6 Why are there different results?

The reason why studies draw different conclusions is because financial performances are evaluated by using financial measures as ROE, ROA, ROC, ROCE, Tobin's Q ratio, etc. that are chosen subjectively and also their method of calculation is arbitrary. The time frame used is different from one study to another and the choice is then indirectly subject to market cycles. The ESG factors valuation of a company is also based on subjective valuation and subjectively are chosen the measures companies decide to release. The reason why different firm in same market could have different reported ESG measures, aside the mandatory ones, is because they could decide to highlight the ones that are more positive than others, and by doing it, giving the investors a distorted view of the overall ESG performances of the firm. Moreover, different legal framework from different countries lead to disagreement in reporting. I will now proceed to elaborate the differences between the European and the United States ESG reporting legal framework.

[7] MSCI obtained the cost of capital by Thomson Reuters. The cost of capital is the weighted average of the cost of equity, debt, calculated after tax, and preferred stock. The cost of equity is calculated from the CAPM by using the equity premium and the risk-free rate of the company's country. The beta is chosen with respect to the country's primary index. The cost of debt includes: for the short-term debt, the one-year yield in the credit curve of the company, while, for the long-term debt is used the ten-year yield on the curve. The cost of the preferred stock has been defined as the current dividend yield on the preferred stock.

1.7 Legal framework

The investor community has been requiring corporation to have ESG-friendly processes and the evaluation of how these processes are conducted is an increasing concern in the due diligence for the investment process. Corporations have been trying to adapt and align their objective with the ones required by their stakeholders but without a legal framework, the internal processes and the reporting could be very different between corporations. Regulators by requiring new ESG organs to be implemented by firms and equal reporting standards can create a standardized framework. This is more important as the attention to ESG factors is increasing and products and processes that are labelled as “green” could in reality have no evidence of being environmentally sustainable [59].

The legal frameworks have still very different objective and requirements with which corporations have to comply. Here are presented the European and US legal frameworks.

1.7.1 European Legal Framework

In 2018 the European commission published the “Action Plan: Financing Sustainable Growth” [60] with the objective of reorienting capital flows to sustainable investing, sponsoring the sustainability aspect into the risk management processes and fostering transparency. They also proposed: The Framework Regulation to facilitate investments that are ESG sustainable, the Sustainable Financial Disclosure Regulation (SFDR) and amendments to the MiFID, UCITS, AIFMD to include the sustainability focus.

1.7.1.1 Framework Regulation

The 18th June 2020 in the Official Journal of EU was published the text for the Level 1 measures[8] [61]. The agreement on the text by the Council and the European Parliament was reached on the 17th December 2019, after the Commission proposed a Regulation on the establishment of a framework to facilitate sustainable investments [62] on the 24th of May 2018.

The framework establishes a classification system (taxonomy) to provide EU member states and corporations the same terminology to identify environmentally sustainable activities. Right now, the sustainability and governance are not included but there is the intention to expand the taxonomy to cover also these subjects.

[8] The European Parliament and Council, in a co-decision process, adopt the laws proposed by the Commission. The level 1 procedure is time consuming and is mostly used for setting out framework principles.

The aim is to reduce climate changes, to build a smooth transition to circular economy, set pollution limits and control, and preserve the biodiversity.

Moreover, Level 2 technical standards[9] will have to be created to further define the concepts that were broadly touched by the Framework Regulation, in particular they will focus on what level of transparency will be required for reporting non-financial information.

1.7.1.2 Sustainable Financial Disclosure Regulation

The SFDR [63] was published on the 9th December 2019 and requires financial market participants[10] to disclose the methods of implementation of measures of sustainability risks in their processes, the disclosure of policies to evaluate and monitor the sustainability of the investments undertaken. Periodic reports will be also created to furnish the impact of the investments. While these measures will involve all asset managers the level 2 measures will give a more specific reference on the disclosure requirements for a specific asset manager class.

1.7.1.3 Amendments to the MiFID, UCITS, AIFMD frameworks

In 2018 the High-Level Expert Group on Sustainable Finance with the report "Financing a Sustainable European Economy" [64] created a set of recommendations in line with the Paris agreement, signed on the 12th December 2015, and the United Nations 2030 Agenda for Sustainable Development Goals (SDGs) to make sustainability a clear guide in the Europe's financial system. On the 8th June 2020, following the "Action Plan on Financing Sustainable Growth", with the raccomadations on the report cited above and taking into account the ESMA[11] final report "integrating sustainability risks and factors in MiFID II" [65], the European Commission published the drafts on the amendements to the Level 2 measures. The four proposed draft are:

- The draft delegated regulation, amending the AIFMD Delegated Regulation (EU) No 231/2013: AIFMs will have to evaluate sustainability risks and adverse impacts on investment decisions.

[9] The Commission, with the consultative bodies, can adopt, adapt and update technical implementing measures. This procedure allows the Council and Parliament to focus on the key political decisions, while technical implementing details can be worked out afterwards by the Commission.

[10] Alternative investment fund managers (AIFMs), Undertakings for Collective Investments in Transferable Securities (UCITS) managers and firms that provide portfolio management.

[11] European Securities and Markets Authority

- The draft delegated directive, amending the UCITS Commission Directive 2010/43/EU: UCITS firms will have to evaluate sustainability risks and adverse impacts on investment decisions.
- The draft delegated regulation, amending the MiFID Commission Delegated Regulation (EU) 2017/565 on organizational requirements. Investment firms will have to evaluate sustainability risks in their risk management processes, they will provide ESG reports under the Article No. 54. There will also be included sustainability risks alongside with costs, risks and complexity in the financial instrument prospects.
- The draft delegated directive, amending the MiFID Commission Delegated Directive (EU) 2017/593). It introduces the theme of "sustainability preferences". Firms will have to take into account sustainability preferences from their clients and will have to evaluate if the financial instrument proposed maintain the initial sustainable characteristics.

The Council and the European Parliament will have to agree on the text of the commission proposal then will be posted in the Official Journal of the EU and the amended rules will come into effect twelve months after the publication.

1.7.2 U.S. Legal Framework

In the United States companies are not required to include ESG disclosures into their Securities and Exchange Commission's (SEC) filings unless the information obtained with the sustainability reports from the company would be in the investors' interests. The SEC's approach to the ESG theme as not satisfied the investor community given the little guidance it requires to corporations, that still use the disclosure of ESG measures as a marketing tool and have difficulties in setting standards to define good or bad sustainability's performances [66].

There have been proposals to set the floor for ESG disclosure requirements but none of them has been approved now. The House Committee on Financial Services proposed ESG disclosure requirements in July 2019 but at the time of current writing has not be approved. In January 2020 the SEC proposed an amendment to the Management Discussion and Analysis (MD&A) rules[12] without including ESG disclosure requirements. Following the proposal, the SEC Commissioner Allison Herren Lee noted

[12] The MD&A is a section of the 10-k (annual) and 10-Q (quarterly) filings in which the management and executives analyze the result from a quantitative and qualitative point of view.

how no requirements were included regarding the ESG factors and urged the SEC to take actions in the future. Moreover, in May 2020 the Investor-as-Owner Subcommittee, guided by the chairman Heidi Stam, of the SEC's Investor Advisory Committee brought the attention to the necessity of disclosure policies on ESG topics. The SEC then announced that is willing to confront with investors and corporations to better understand how they uses and evaluate ESG information and to create an adequate disclosure framework. The procrastination of taking measures risks to set U.S. companies at a competitive disadvantage.

Data from the Governance & Accountability Institute shows that in 2011 of all the companies in the S&P 500 only 20 percent released sustainability reports, in 2015 81 percent and in 2019 nine out of ten released an ESG report.

This trend has been sponsored by the United Nations Global Compact (UNGC) that now has more than 12'000 companies in 160 countries around the world that underwrote labor, environment and human rights principles. The United Nations Principles for Responsible Investing (UNPRI) encourage investors to incorporate ESG factors when making investment decisions.

1.7.2.1 Investor proactivity

While the common barrier to ESG integration for institutional investors have been believed to be lower returns, Eccles et al. [67] conducted a study on 582 institutional investors, across America, Asia Pacific, Europe, Africa, Middle East and found that the major drawback is the lack of high-quality data to correctly evaluate companies.

Since the legislation in the United States is lagging, some of the biggest institutional investors such as Black Rock, State Street and Vanguard, which hold more than 20 percent in the companies in the Standard & Poor 500, are requiring companies to disclose their ESG measures with the SASB or the TCFD framework [68].

The SASB framework provides guidance on the disclosure of ESG metrics, in particular: air pollutants[13], water management[14], greenhouse gas (GHG) emissions[15] (calculated according to the EPA Renewable Fuel Standard 2 (RFS2) requirements), operational

[13] NOx (excluding N2O), SOx, volatile organic compounds (VOCs), particulate matter (PM), and hazardous air pollutants (HAPs).

[14] total water withdrawn, total water consumed, water quality and strategies to manage these measures.

[15] in grams of CO2-e per megajoule.

safety[16], emergency preparedness, environmental impacts, data security. The guidelines for the SASB framework are sector-specific, while for the TCFD framework are general and sector-specific.

The TCFD framework only provide guidelines on climate related topics, as how to manage resources, pollution and usage of alternative energy in the production processes. The framework [69] has been adopted by regulators in EU and U.K..

1.8 A sustainable CAPM approach

The Capital Asset Pricing Model (CAPM) of W. Sharpe (1964) [70] and Lintner (1965) [71] is based on the mean-variance model developed by Markowitz in 1959. In the portfolio choice theory, the investor decides where to allocate his capital on a pool of assets that maximize the expected return given the variance and that minimize the variance given an expected return. The expected return of a portfolio ($\mathrm{E}(r_P)$) is given by the average of the single asset returns while the computation of the portfolio variance (σ_P^2) include the correlation between the assets that could give the diversification effect:

$$\mathrm{E}(r_P) = \sum_{a=1}^{n} w_a \, E(r_a)$$

$$\sigma_P^2 = \sum_{a=1}^{n} \sum_{i=1}^{n} w_a \, w_i \, \sigma_{ai} = \sum_{a=1}^{n} w_a^2 \, \sigma_a^2 + \sum_{a=1}^{n} \sum_{i=1}^{n} w_a \, w_i \, \sigma_{ai}$$

Sharpe and Lintner added to the Markowitz theory the assumptions of investor agreement on the joint distribution of asset returns in a one period time and the assumption of borrowing and lending at the risk-free rate (r_f) independently from the quantity selected.

The efficient portfolios are all the ones on the line that connects the risk-free rate and the tangency portfolio (T) that lies on the efficient frontier. The efficient frontier is the curve formed by the combination of assets that minimize the return variance for different levels of expected returns. The portfolios on the frontier do not include the borrowing and lending possibilities.

[16] The accounting metrics are: Process Safety Incidents Count (PSIC), Process Safety Total Incident Rate (PSTIR), and Process Safety Incident Severity Rate (PSISR).

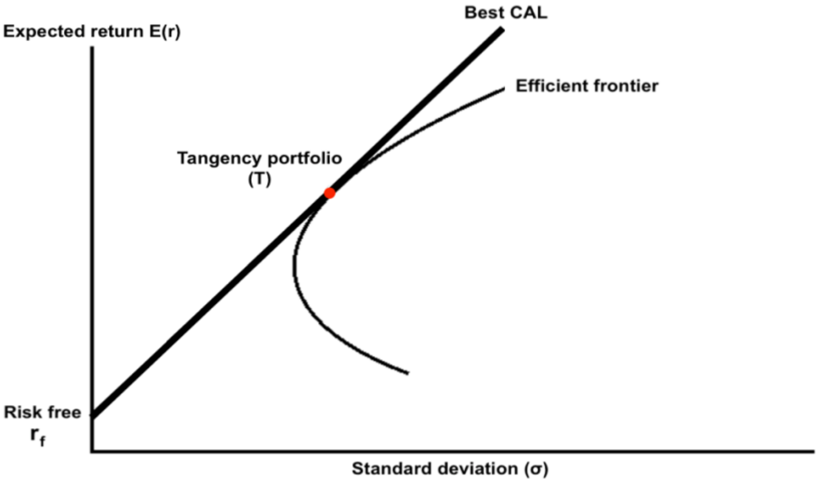


Figure 1.1: Efficient frontier

Tobin's separation theorem [72] shows how the selection of the efficient portfolio created by the combination of the risk-free rate for borrowing/lending and the tangency portfolio is different from each investor since risk-aversion is subjective. The efficient portfolios are the ones that lie on the Capital Asset Line (CAL). The slope of the CAL is the reward to variability ratio [73], later called the Sharpe ratio and it is used to measure the performance of a portfolio. The formula is:

$$Sharpe\ ratio = \frac{r_P - r_f}{\sigma_P}$$

The CAPM relates the expected return of an individual asset and its risk factor represented by the beta in the following formula:

$$E(r_i) = r_f + \beta_i[E(r_m) - r_f]$$

The first term is the risk-free rate (r_f), that is the return expected by an asset that has zero systematic risk and therefore a beta equal to zero. The beta (β_i) of the i-asset is the sensitivity of the asset returns to the market returns, it is the regression of the asset return on the market return. If instead of a single asset we would have had an entire portfolio, and in particular the market portfolio, the risk of the portfolio, measured by the variance of its return, is a weighted average of all the covariance risks of all the assets included in the portfolio. The relation for the single asset is shown in the beta formula below:

$$\beta_i = \frac{cov(r_i, r_m)}{Var(r_m)}$$

The beta of the i-asset in the market portfolio is the covariance risk of the i-asset relative to the average covariance risk of all the assets (i.e. the variance of the market return). This is demonstrated by:

$$\sigma^2(r_m) = Cov(r_m, r_m) = Cov(\sum_{a=1}^{n} w_a r_a, r_m) = \sum_{a=1}^{n} w_a Cov(r_a, r_m)$$

The w_a is the weight of the i-asset in the market portfolio.

The beta is multiplied by the market premium, that is the difference between the expected market return and the return from a risk-free asset. This difference is the slope of the security market line (SML) that represent the relation between the expected return and the systematic risk.

The observed returns have not been fully explained by the CAPM model and new models were developed. In particular the Three Factor Model by Fama and French [74] with which they explained the cross-sectional variation in average stock returns, in particular they added the firm size factor (small minus big) that accounts for the excess return achieved by small versus high market capitalization companies; the value factor (high minus low) that accounts for higher returns from value stocks versus growth stocks and lastly the excess return on the market. The Fama French model is:

$$E(r_i) - r_f = +\beta_1[E(r_m) - r_f] + \beta_2 E(SMB) + \beta_3 E(HML)$$

Zerbib [75] proposed a sustainable CAPM with partial segmentation and based on heterogeneous preferences among investors. He found that two taste premia and two exclusion premia describe the excess returns that were not predicted by the capital asset pricing model. There are 3 type of investors: the regular investors, the investors that perform exclusionary screening and the ESG investors that choose the stocks valuing their sustainability. The regular and the excluder investors have an exponential utility function:

$$\max_{w_j} \mathrm{E}\left(U_j\left(W_{j,t+1}\right)\right) = \max_{w_j} \mathrm{E}\left(1 - e^{-\psi_j^a W_{j,t+1}}\right)$$

These two types of investors select the weights of assets based on the equation above. ψ_j^a is the risk aversion of investor j, while $W_{j,t+1}$ is his wealth level at time *t+1*.

Integrators (i.e. ESG investors) adjust their function by including deterministic costs of externalities [76] denoted here by the vector C^W, and has the same unit of wealth. These investors decide the weights of the risky assets based on:

$$\max_{w_i} \mathrm{E}\left(U_i\left(W_{i,t+1}\right)\right) = \max_{w_i} \mathrm{E}\left(1 - e^{-\psi_i^a W_{i,t+1} + w_i' C^W}\right)$$

From the equation we can see how the utility decrease as the costs increase.

Every investor act in a one period time, in which they want to increase their wealth, the increase is evaluated only from the amount they have at time *t*, excluding other sources of income that could occur from *t* to *t+1*. The returns are assumed to be gaussian distributed. The market is perfect[17], borrowing and lending have no constrain and are performed at the same interest rate. In the model the Short selling is allowed. The I_n assets are the investable assets, while the X_n assets are the excluded assets. The proportion of the market value of the excluded assets is $q \in [0,1]$. The wealth shares of excluders is p_e , for integrators is p_i and for regular investors is $1 - p_e - p_i$. Moreover ψ is defined as the relative risk aversion of all investors (or global risk aversion) that is a measure representing the risk preferences of a decision maker based on the possible outcomes that are the arguments of the utility function [77]. Also, c_k is the cost of externalities for the k-stock, this cost is present when investors have preferences between the stocks. c_{m_I} is the cost of externalities of the value-weighted portfolio of the investable (I_n) assets. The vector of costs for a ϑ-class of assets is $C^\vartheta = (c_{\vartheta_1}, \dots , c_{\vartheta_{n_\vartheta}})'$. In the model proposed by Zerbib the time subscripts are omitted and the returns are the excess returns on the risk-free rate.

The expected excess return on an asset $k \in \{I_1, \dots, I_{n_I}, X_1, \dots, X_{n_X}\}$ from the S-CAPM model is:

$$\mathrm{E}(r_k) = \beta_{km_I}\left(\mathrm{E}\left(r_{m_I}\right) - p_i c_{m_I}\right) + \frac{p_i}{1 - p_e} c_k - \frac{p_i p_e}{1 - p_e} \beta_{kI} C_I$$
$$+ \psi \frac{p_e}{1 - p_e} q Cov\left(r_k, r_{m_X} \middle| r_I\right) + \psi q Cov\left(r_k, r_{m_X} \middle| r_{m_I}\right)$$

Exclusion and taste premia have cross-effects on investable and excluded assets. $\frac{p_i}{1-p_e} c_k - \frac{p_i p_e}{1-p_e} \beta_{kI} C_I$ is the taste premia, while $+\psi \frac{p_e}{1-p_e} q Cov\left(r_k, r_{m_X} \middle| r_I\right) + \psi q Cov\left(r_k, r_{m_X} \middle| r_I\right)$ is the exclusion premia.

In particular the excess return for I_k, $k \in \{1, \dots, n_i\}$ is:

$$\mathrm{E}\left(r_{I_k}\right) = \beta_{I_k m_I}\left(\mathrm{E}\left(r_{m_I}\right) - p_i c_{m_I}\right) + p_i c_{I_k} + \psi q Cov\left(r_{I_k}, r_{m_X} \middle| r_{m_I}\right)$$

[17] A market where the two parts, a buyer and a seller, have complete information, there is no monopoly and prices are not manipulated.

The $p_i c_{I_k}$ is the direct taste premium, while $\psi q Cov\left(r_{I_k}, r_{m_X} \middle| r_{m_I}\right)$ is the exclusion market premium.

The excess return for X_k , $k \in \{1, \dots, n_X\}$ is:

$$\mathrm{E}\left(r_{X_k}\right) = \beta_{X_k m_I}\left(\mathrm{E}\left(r_{m_I}\right) - p_i c_{m_I}\right) + \frac{p_i}{1-p_e} c_{X_k} - \frac{p_i p_e}{1-p_e} \beta_{X_k I} C_I + \psi \frac{p_e}{1-p_e} q Cov\left(r_{X_k}, r_{m_X} \middle| r_I\right) + \psi q Cov\left(r_{X_k}, r_{m_X} \middle| r_{m_I}\right)$$

In particular: $\frac{p_i}{1-p_e} c_{X_k}$ is the direct taste premium, $\frac{p_i p_e}{1-p_e} \beta_{X_k I} C_I$ is the indirect taste premium, $\psi \frac{p_e}{1-p_e} q Cov\left(r_{X_k}, r_{m_X} \middle| r_I\right)$ is the exclusion asset premium and $\psi q Cov\left(r_{X_k}, r_{m_X} \middle| r_I\right)$ is the exclusion market premium. The presence of the exclusion market premium on the formula for the excess return of the I_k investable assets and the presence of the indirect taste premium in the excess return formula from X_k excluded assets show a cross relation between an ESG inclusion and an exclusion screening investment strategy.

Moreover, as shown by the formula of the direct taste premia, $p_i c_{I_k}$ *and* $\frac{p_i}{1-p_e} c_{X_k}$, for investable ($I_k$) and excluded ($X_k$) assets, respectively, increase as the cost of externalities (c_{I_k} *and* c_{X_k}) increases, this is because the integrators have to have a form of compensation and incentive, proportional with the costs of externalities, for the acquisition of a specific asset. This is in line with Pastor et al. [70] findings that stated that brown assets[18], on average, have positive alphas, while green assets have negative.

1.8.1 Three different preference cases

The first one is the one where no asset is excluded so $p_e = 0$, but investors have different assets taste $p_i > 0$. Given this scenario the exclusion premia is equal to zero, in particular $q = 0$. The market coincides with the investable market ($m_I = m$). The expected excess return on the k-asset is:

$$\mathrm{E}(r_k) = \beta_{km}(\mathrm{E}(r_m) - p_i c_m) + p_i c_k$$

The second limit case showed by Zerbib [75] is when ESG investors have no preferences on assets with respect to the other investors ($c_k = 0, \forall k \in \left\{I_1, \dots, I_{n_I}, X_1, \dots, X_{n_X}\right\}$).

[18] Brown assets are “not climate proofed”, companies do not perform an active focus on climate impacts.

But they exclude some type of non ESG-friendly assets ($p_e > 0$, $p_i = 0$), so the taste premia part of the equation, $\frac{p_i}{1-p_e}c_k - \frac{p_i p_e}{1-p_e}\beta_{kI}C_I$, is equal to zero. Then the equation of the excess returns of investable assets is:

$$\mathrm{E}\left(r_{I_k}\right) = \beta_{I_k m_I}\mathrm{E}(r_m) + \psi q Cov\left(r_{I_k}, r_{m_X} \middle| r_{m_I}\right)$$

Third and last case is when there are not sustainable investors, that is $p_e = 0, p_i = 0$ and the excluded assets are zero: $q = 0$, then the market coincides with the investable market: $m_I = m$ and the model is now reduced to the CAPM.

The expected excess return of excluded assets with respect to the total asset in the investable market is:

$$\mathrm{E}\left(r_{X_k}\right) = \beta_{X_k m_I}\mathrm{E}\left(r_{m_I}\right) + \psi \frac{p_e}{1-p_e} q Cov\left(r_{X_k}, r_{m_X} \middle| r_I\right) + \psi q Cov\left(r_{X_k}, r_{m_X} \middle| r_{m_I}\right)$$

2 Chapter 2
Web Scraping

2.1 Introduction

Websites have a very large amount of invaluable data, and it requires techniques to access the information. Instead of copying the information manually into a new document, web scraping is a method that was implemented to access the required information from a group of the websites automatically [78]. This technique extracts the data from the World Wide Web and process it into simpler structure by saving it to databases, spreadsheets or CSV file to then perform analysis. The web is composed by unstructured and structured data, in different formats and from different sources. Based on the task performed the amount of useful information from the data can vary. Web scraping is a time-consuming, resource-consuming, complex task given the aforementioned differences the process has to deal with. This complexity increases as the number of analyzed sites increases.

The web pages can be scraped manually by a user, by human aided procedures, or automatically through the use of a crawler that parses different links from the domain page and extracts the data. The method used depends on the number of tasks the bot has to perform to retrieve the data. If the page scraped is one, most probably an ad hoc code will better suit the job and will be more efficient and less time consuming than creating a more generic scraper tool that would require further data manipulation of the retrieved data for the following analysis.

Other than collecting data from parsing markup languages[19] or JSON files, web scraping tools can perform data visualization and use natural language processing to perform the task as a human would do. With the increasing requirements a bot has to perform there is the risk of having a very specific code, not useful for the broader scope of targeting different web sites. To compensate this drawback the complexity will have to rise.

[19] A system to annotate a document with a significant difference from the text. When the text is shown the markup language is not displayed. An example is HTML that has also presentation semantics to specify how structured data are presented on a media.

Given the vast amount of data being present in the WWW, the web scraping technique has been believed to be an efficient and powerful tool to collect useful information from websites [79] [80].

2.2 The Web Scraping process

The process of web scraping is made of crawling the target web page and then scraping it. To scrape the data, that is collecting the data, the program has to acquire the data from the target site first, this action is called crawling. In particular the bot is opened and a Hyper-Text Transfer Protocol (HTTP) request is made using a GET query from a URL[20]. The GET request retrieves the data from the server. GET is only one of the multiple requests that can be performed, examples are POST and DELETE, used respectively to upload and delete from a server.

The most common modules used for requesting other than GET are Urllib2 and selenium. To complete the task a set of functions such as authentication, redirection, cookies enablers or deniers are used by the programs created. Most of the time there is no need of custom settings but depends on the complexity of the task and in particular of the site to which the request is made. The code in my thesis uses the default HTTP method GET, but with Scrapy (i.e. the framework used to web scraping in this thesis) the request customization can be done with specification of the method() parameter. Sometimes the request won't be able to get to the server of the site because limitations for bots are set in place. In particular, there are two very important aspect to consider, one is the user agent, that is the identification number of the software user that acts on behalf of a user in a server, and the other one is the robots.txt page of the site. With regards to the first one there is the risk that the identification number of the bot has been compromised and restricted/banned from operating/accessing the server; the second is a web page, specific for every company, that sets limitations on the bot when dealing with information on the site. In scrapy changes of the settings code lines of the spider (i.e. the class that describe how the bot operates) will allow to eliminate specific constrains that can affect the retrieving process.

After getting the response, the previous request is downloaded and the extraction process begins. Scrapy has its own mechanisms, called "selectors", because the coder is able to select through them the specific text parts of interest. I decided to code the XPath selector

[20] Uniform Resource Locator

to extract the data but, as I will show later, I have not created an ad-hoc, specific code for each analyzed site since it would have been inefficient and time consuming. The unstructured data that I have retrieved, given this decision, will be cleaned with another program I wrote and that will be discussed on the “Data Cleaning” section.

Now that the data are extracted, with another function, called “callback”, the web scraping program will process the data received and return the scraped data and/or the URLs connected to the scraped pages to be read by the user or to be reprocessed to perform subsequent tasks. The callback function will redirect the data to another function that based on the script will perform other tasks.

The data can be retrieved with more structure by writing more specific selector codes, this is particularly useful when the scraped pages have the same structure, the same objects, the same paths and the same markups. An example are the Amazon.com webpages that have a standardized format. In the Scrapy framework a more structured data storing, as creating data classes, subclasses and other classifications, can be done by retrieving the data as items with the item() function.

Given the different sites I have to analyze and their differences in form and structure I found a good tradeoff between complexity of the codes created for each company and the ability of the bot of retrieving the data in a structured form. Most of the developed codes save the text in an unstructured form as shown below:

```
<div class="text-content">La rendicontazione non finanziaria riflette il principio di materialit√† o
rilevanza. L?*elemento…
```

It includes markups (“<div class="text-content">”) that would affect the counting method performances to establish the relevance of sustainability words concentration in the text analyzed, given that “div”, “class” and other markups would be included in the total word count. From the elevated presence in the text, these words or acronyms would have been the most recurring ones in nearly all the text files analyzed.

From the text above we can see how the accents and apostrophes are not read by the spider, moreover, the symbols that represent them change as the previous and the following letters change. This problem will be addressed when creating the dictionary that will be used to select and count the words that have a sustainability connotation.

The more the code is non-specific the better it is, given also the differences between the web sites analyzed. Unfortunately, all the sites present different form and structure therefore, I have created a web scraping program that allows for a good trade-off between the specificity of the code and the loss of information that non-specific codes could have.

2.3 Web scraping techniques

The exchange of information between people and computers on the internet happens through protocols. These protocols set the rules on how the information is passed between the agents, what happens when information does not arrive to the recipient, how a request to obtain specific data has to be done, and many other more. To set a standard of communication the International Organization for Standardization (ISO) created the model ISO/OSI that divides the process of communication into seven layers on which specific tasks have to be processed in order to maintain trustworthiness, security and integrity in the process. The most known protocols are HTTP and HTTPS and are used to recall websites.

One of the first protocol on the internet is the File Transfer Protocol (FTP), defined in 1985 by J. Postel and J.K. Reynolds [81], this protocol, as the name says, was created to define the upload and download of entire documents, files between personal devices and servers. Programmers started to create automated programs that were able to retrieve and collect data on the internet. Bots and web crawlers were born [82].

One of the main problems in retrieving data is that there is no download button for the webpages, and even back then, one web scraping technique, very tedious and inefficient, was to manually select the parts of interest to copy and paste them on a personal computer. Then programmers developed more advanced, automated, techniques to extract data. These, in particular, are divided into three approaches: supervised, unsupervised and hybrid [83]. The supervised approach uses Machine Learning (ML) with learning models to predict the extracting patterns from a given data set. The unsupervised approach aims to eliminate the manual labelling process[21] by automating the process of identify the targeted data to extract. This method can be further divided into three groups: template detection, statistical-based, and assumption techniques. The first one is a method that identify the main block of interest, i.e. menus, headers, footers, ads, content sections and others, and retrieves it after evaluating all the blocks on the page [84]. The Statistical

[21] It is the process of manually detect and tag specific samples of data. It is important to classify data especially for ML where the data set is the base on which the computer is trained. Recognize specific patterns is key to identify for example if a group of words are referred to a specific semantic group.

method is the one that uses parameters as the link density, the tag ratio, the number of nodes, etc. to detect the targeted part of the page the programmer needs. The assumption technique creates different learning processes based on assumptions that are made on the structure of the site.

Lastly there is the hybrid approach that is an improvement of both approaches: supervised and unsupervised [85]. The first step of this approach is to divide the webpage analyzed with a Document Object Model (DOM) structure and then extract the text and the information with a statistical approach rather than ML repeating training and manual labelling [86].

The different approaches differ especially in the time cost of the task. Since in my research the time is not a variable that affect the outcome of the process, I decided to not use the unsupervised and the hybrid approach.

2.4 Scrapy Framework

Scrapy, that is written in Python, is an open-source web crawling framework. It was released in 2008 and it is used to crawl and scrape structured data from websites. This framework allows to control the data flow, and in particular, through its crawling and scraping activities it is able to fetch the web pages, send requests to the sites' servers and with its custom classes parse responses and extract items.

I decided to use this framework given the big user base and the relatively fast learning curve.

Scrapy also is very useful to accelerate the creation of crawling and scraping projects given the standardized files it creates to run the bots on the targeted websites. This framework also allows the use of a shell that fetches the targeted site and simulates the behaviors of human users that is useful in the writing phase of the code.

There are also other non-trivial problems when creating a web scraping code given that in the recent years the attention of programmers has been on the visual rendering of web pages to create a more friendly experience for the user, that consequently led to a more complex structure of the site. Asynchronous Java Requests (AJAX) that allow dynamic loading of web pages as the users scroll down or CSS to separate the style from the contents are an example. These improvements on the user end make the extracting process harder and increase the error numbers on the applications and browsers' extensions that do not require coding.

I decided to use Scrapy also to not incur in these errors that would not be easily addressable. I created a specific code for each company analyzed to have a high degree of freedom on the extraction code characteristics.

2.4.1 Spiders

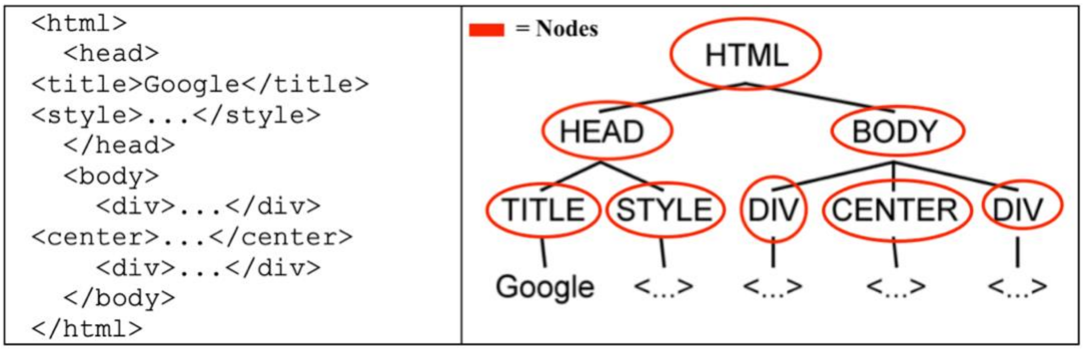
Spiders are classes that the programmer has to modify in order to define what are the targeted sites, how they will be scraped, how the crawling process will be performed and how the extraction process will work.

The process begins with an initial request to the starting URL that the programmer has to specify in the “start_urls”. Then it has to be specified a parse method with a callback function in order to process the downloaded response (webpages) from the request. The parsing action of the webpages’ contents is done by using selectors, XPaths in our case. Also, the request in the parsing function contains a callback that will then be downloaded and processed. Lastly, the item objects are generated from the parsed data and stored in a database (by writing the Item Pipelines) or written in a JSON, CSV, XML file (by setting Item Exporters or using Feed exports when running the code from the terminal).

2.4.2 Xpaths

The Scrapy framework allows the writing of expressions to select and extract the text from HTML or XML sources by using XPath expressions or other methods and libraries such as BeautifulSoup or lxml with the extended Cascade Style Sheet (CSS) selectors.

They are a pattern of elements that suggest the browser which HTML elements, called subjects, the bot should select. In particular, given the tree structure of websites, the XPath language is used to navigate the tree and select the nodes (see figure 2.1) in structured data extraction algorithms [87].



```
<html>
  <head>
<title>Google</title>
<style>...</style>
  </head>
  <body>
    <div>...</div>
<center>...</center>
    <div>...</div>
  </body>
</html>
```

Fig. 2.1 HTML code on the left side and its tree structure on the right with nodes

An XPath selects following nodes, starting from a root node (“html” in fig.2.1). These selectors are expressed by a syntax and can be reduced by using abbreviations [88]. The XPath can be written in two methods: relative and absolute location. The first one is one location step, or a sequence of location steps, / separated, that divide the subsequential nodes. The second one, absolute XPath, is a sequence of / separators that starts selecting the root, followed by relative XPath locations that select the nodes where the targeted information is. In the example above (Fig. 2.1) if the user wants the bot to retrieve “Google” can use the following methods:

```
Relative XPath:          //title
Absolute XPaths:         /html/head/title
```

Fig. 2.2 Example of Relative and Absolute XPaths

There is evidence [89] that the use of a relative path to extract the data can improve robustness, that is the ability to deal with inputs’ errors and execution errors. This is an important future and it is particularly in my project since, by analyzing different web sites’ domains and different web pages without a common tree structure, I had to retrieve the data without writing the paths from the root to the context node (i.e. absolute XPaths) since they would have differed from one project to another as the trees, from one page to another, are different.

In particular, if a targeted site has the following HTML code:

```
<html>
   <head>
<title>Google</title>
<style>…</style>
   </head>
   <body>
      <div>
        <title>
          <p> “the sky is “
             <strong>blue</strong>
          </p>
        </title>
     </div>
   </body>
</html>
```

Fig.2.3 Example of a website code

The Absolute XPath (fig. 2.2) would not extract both “Google” and “the sky is blue” but only the first one. To also extract the second text a new Absolute XPath locator must be created. On the other side, with the Relative XPath (fig. 2.2) the extracted results would have been:

"Google" and "<p> "the sky is" <strong>blue</strong></p>". Can be noted that the text with this method is completely retrieved but in an unstructured form.

In the bots writing process I decided to use the relative path and then process the text afterwards by eliminating the markups.

2.4.3 Crawling

The method to retrieve the pages used above is related to a single web page. For this project is important to retrieve all the pages that match specific characteristics and then extract their data with the method seen in the previous paragraph.

Crawling is the action to collect all the web pages starting from a seed URL from which the crawl starts extracting following URLs. In the Scrapy framework, inside the spider folder, in the spider code, there is a code line called "start_urls" where is written the URL from which the bot starts the extracting and crawling processes. Scrapy checks if the web pages are under the specified domains in the "allowed_domains" list, if the response is true the bot starts to execute the scraping task. In particular for the crawling action the object is "linkextractor" which defines the processes and the methods to process the web pages that are crawled.

The method used to scrape crawled URLs is first-in-first-out, this strategy is called breadth-first. The first link extracted is processed and then the bot goes on. If a link matches URLs already analyzed, parsed and scraped the bot will not re-process it but will pass to the following one to avoid double processing. The following image (fig. 2.4) resumes the process.

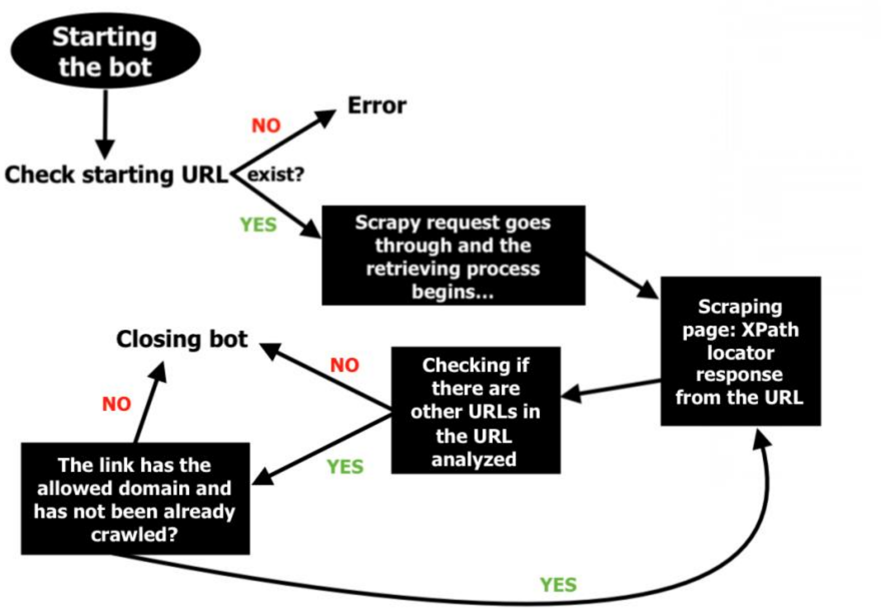
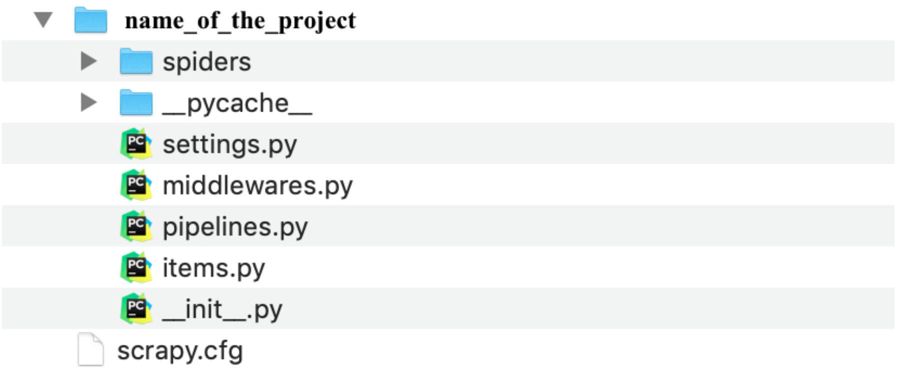

Starting the bot
Check starting URL
exist?
NO
Error
YES
Scrapy request goes through and the retrieving process begins...
Scraping page: XPath locator response from the URL
Checking if there are other URLs in the URL analyzed
NO
Closing bot
YES
The link has the allowed domain and has not been already crawled?
NO
YES

Fig.2.4 Web scraping bot process

2.4.4 Creating the web scraping code

Starting with opening the terminal it has to be selected the directory where the code will be stored. Then with “scrapy startproject name_of_the_project” a folder with the name chosen is created. Inside there are a “__pycache__.py”, a “spider” folder and the following files: “setting.py”, “middlewares.py”, “pipelines.py”, “items.py”, “__init__.py”, “scrapy.cfg”. In fig. 2.4 there is a representation of the folder created.

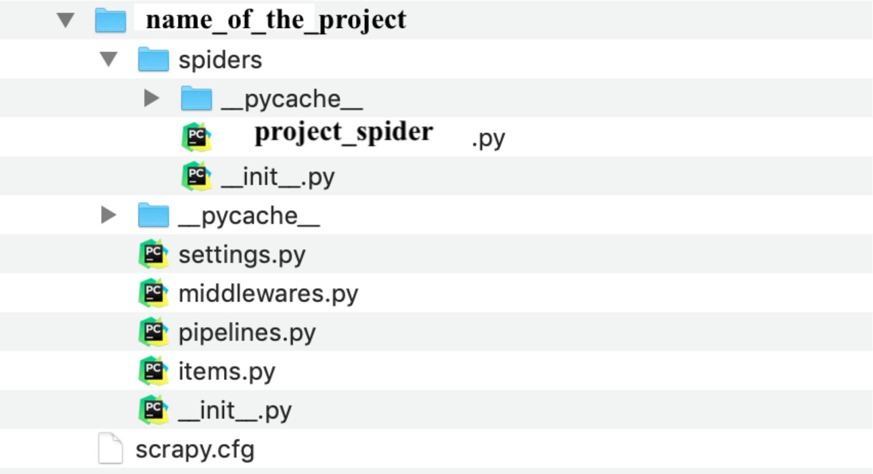


```
▼ name_of_the_project
    ▶ spiders
    ▶ __pycache__
      settings.py
      middlewares.py
      pipelines.py
      items.py
      __init__.py
  scrapy.cfg
```

Fig.2.4 Scrapy created folder

After creating the project folder, a spider has to be created in order to perform web scraping tasks.

From the terminal, and after selecting the name_of_the_project directory, is created a spider with the following code line “scrapy genspider project_spider example.com”. The spider, called project_spider, and a new _init_.py file are now positioned under the spiders folder (fig. 2.5).

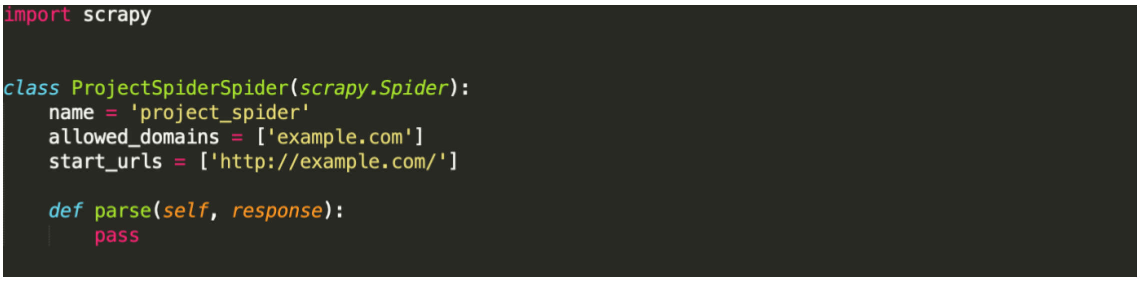


```
▼ name_of_the_project
    ▼ spiders
        ▶ __pycache__
          project_spider  .py
          __init__.py
    ▶ __pycache__
      settings.py
      middlewares.py
      pipelines.py
      items.py
      __init__.py
  scrapy.cfg
```

Fig.2.5 Scrapy project folder after creation of the spider

Inside the "spiders" folder there is the "__pycache__" directory that is created when a program in Python is run. The interpreter, that is a program that reads, execute the written code and compiles the code to bytecode (also called portable code). The bytecode is an optimized version of the program that the programmer has created, it gives instructions to run the code faster. As the code is modified, the bytecode will change accordingly. If the "__pycache__" file is delated by the programmer it will be recreated at the next run, unless specified otherwise. Practically, for the intent of this thesis, this directory is ignored as the presence of it doesn't affect the processes.

The project_spider.py file is the actual spider that the programmer has to modify in order to perform the required tasks. The spider created with genspider, from the example above, is presented in fig.2.6.; while the spiders used to analyze the sample of companies can be found in Appendix A.



```python
import scrapy

class ProjectSpiderSpider(scrapy.Spider):
    name = 'project_spider'
    allowed_domains = ['example.com']
    start_urls = ['http://example.com/']

    def parse(self, response):
        pass
```

Fig.2.6 Initial spider code

The "__init__.py" file is a constructor in python classes, and it allows the creation of attributes in a Python class creation process. These specifications are not needed for the purpose of this project and therefore the file is empty.

The settings.py file allows the customization of the characteristics of the execution components of Scrapy. In particular by changing the settins.py file the user can affect the core, the extensions, the pipeline and the spider.

Middlewares.py is the code needed to post process the inputs, output and exceptions. It acts in between of the engine and the spiders to process responses (inputs) and requests (outputs).

Pipelines.py is the file that processes the items, outputs, of the webpages scraped by the spider. In particular it could validate the data, by checking if they contain fields, drop double counted items, store the results of the spider into a database.

Items.py is used to define the fields of the items scraped, for the purpose of the project it is not needed to specify the type of the items that are extracted. This file as well as pipelines.py, middlewares.py and settings.py are shown in Appendix A.

Scrapy.cfg is a file containing specific information about the Scrapy project created. It contains the name of the Python module and its settings for deploying purpose. It is stored in the project root directory.

2.5 Legal Framework

Web scraping is not only used by third parties that wants to extract data from targeted sites but also companies can decide to analyze their own site to check if their targets are in line with the structure and the content of the site. Web scraping methods can be used to act in different scenarios to scrape different topics such as contacts, prices, product reviews, weather data and many more. Although web scraping is a useful technique to collect data some controversy arose and still arise especially on copyrights [90] and terms of services (ToS) [91]. Nowadays the major concerns come from the scraping of social networks that contain a massive amount of sensible data. Before the Facebook-Cambridge Analytica data scandal there was not much attention on this issue, given also the difficulties on demonstrating any infringement of policies.

In particular in the U.S. the US CFAA law[22] of 1986 does not apply to web scrapers (if the server is not violated) since users only reach the websites' publicly available information without accessing any prohibited area that is not present on the public website [92]. In some jurisdiction under the CFAA merely breaching a website’s terms of use can

[22] Federal criminal law that makes unlawful computer-related activities involving the unauthorized access of prohibited information.

potentially expose a web scraper to liability. Even if in 2017 the District Court for the Northern District of California ruled favorably for web scrapers that extracted publicly available information, the uncertainty on what can be extracted from a website still remains [93].

In the European Union there is the General Data Protection Regulations (GDPR) that sets specific rules for data protection when it comes to data gathering. Web scraping legislation varies by location but follows common rules as attention to learning about good web citizenship, about not violating copyright, not violating ToS and paying attention to robots.txt rules (a document that sets limits of content that can be accessed by web crawlers). Even if there is no law about obeying to robots.txt files if the site owner could demonstrate lower site's performances there could be some repercussion on the bot user.

3 Chapter 3
ESG valuation of Italian companies

3.1 Introduction

The web scraping ESG valuation is based on 32 companies of which 17 are in the Energy sector, the remaining 15 are the top companies by market cap, as of the 10th of February 2021, by Borsa di Milano. Each site is analyzed in order to retrieve the most coherent texts related to the sustainability topics. External links are excluded as well as all the text files such as sustainability reports.

After the bot retrieves all the text in an unstructured form and stores it in a csv file a new code reprocesses it and counts the words in order to perform the analysis to evaluate the relation between my assigned values and the Bloomberg data, extracted from the Bloomberg Terminal.

The explanation of the project starts by presenting the writing process of the specific companies' codes. Then it is explained the data cleaning code in order to eliminate all the markups in the retrieved text. The total word frequency is shown for each company.

It is then created a dictionary of words, grouped based on their E, S, G semantic fields, that is compared to the total of individual website's words. By calculating the recurrency of the chosen words on the overall pool of words, the project E, S, G, company specific scores are determined. The frequency of the dictionary's words is plotted for each company. Lastly, a linear regression analysis and a rank regression on the Bloomberg ESG scores are performed and the conclusions are drawn.

3.2 Writing the Python crawler and scraper

The code created has not been ad-hoc for every analyzed company's site since it would have been time consuming and with little efficiency to my scope. While a non-specific approach has been chosen, I had to check each site and apport modifications to the individual codes based on the structure and the complexity of the web pages. A specific code that retrieves all the words in the sustainability pages for each company analyzed has been written.

In fig. 3.1 and fig. 3.2 are presented the two spiders created on the specifics of the targeted website. The first image is the code of a web crawler, web scraping code, while the second is a scraping code of specific targeted URLs.

It is now modified the code in fig 2.6 that was already automatically created with the specific project name and website. In particular, in the first image the first five rows are the code to import the modules that will be used in the composition of the class[23]. The name of the class is automatically created from the name of the spider that has been given when creating the initial folders. Then it is modified the parent class of the class to "CrawlSpider" since a crawling process will be performed. In fig.3.2 on the other hand, the base class is "scrapy.Spider" as it is in fig.2.6 since no web crawling step is performed. The name, the allowed domain and the start_urls variables are created automatically. These variables can be modified, in particular in the "start_urls" the "www" in the URLs has to be eliminated to have better performances.

The rules variable is then created to set the mechanisms that the code has to perform when a new URL is crawled. The first "Rule" object is "LinkExtractor" that contains a parameter, i.e. "allow", that defines an expression that a crawled absolute URL has to match in order to be extracted. In the example in fig. 3.1 it is "sostenibilita", sustainability in English, so every time is found a link that contains this word and has not been already retrieved, the bot processes the page. There are also other parameters that can be specified as "deny()", "allow_domains()" and many more that are used in order to specify the characteristics of the targeted URLs.

The "callback" variable is set as the function that will process each link extracted after is crawled. The last parameter of the LinkExtractor object is "follow" and is a Boolean variable that if set to "True" will allow a subsequent extraction of links from the extracted links, while if set to "False" only the links extracted from the starting URLs will be processed and then the code will stop running.

The "parse_text" function defines the tasks the bot will have to perform in order to extract the text from the strating_urls and the extracted links.

[23] A class is an object constructor, a template to create objects. While in a function, e.g. def nome(x,y):, the x, y and eventually others letter/words are the parameters of the function, on "class nome(x)" there are not specified any parameters since the "x" is the base class. This base class, called also parent class, is the class from which the "nome" inherits all the properties and methods.

In this particular example the variable “text” retrieves all the text under the “p” tag of the targeted site. To check the tags under which the text is stored the programmer has to manually check them by inspecting the targeted site pages that are going to be scraped.

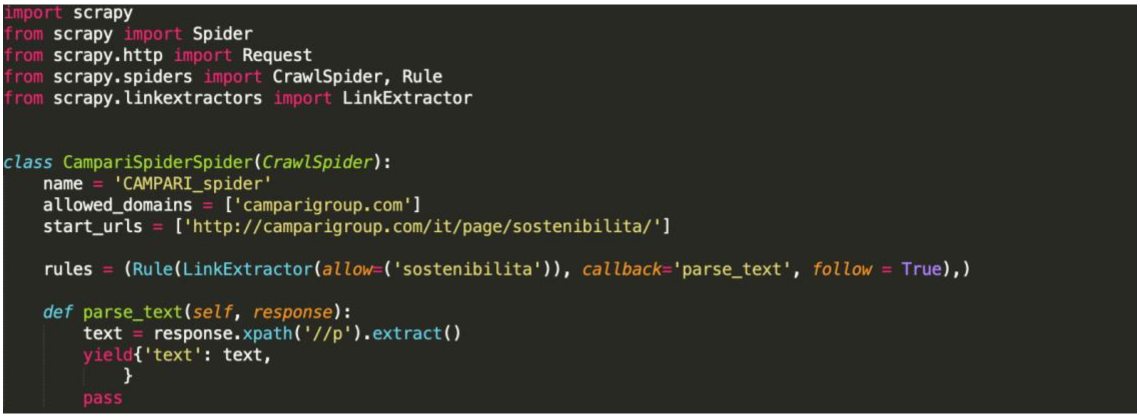
A selector, “xpath”, is wrapped over a response that is a HtmlResponse / XmlResponse[24] object that is used to select and extract data.

If there is more text to be retrieved under different tags or that perhaps it is shown gradually given the dynamic loading process of the pages analyzed, new variables have to be created (Appendix A).

Lastly the text variables are yielded. I used the yield function since, in contrast to “print”, it allows the storage of data in a database or in text files such as CSV, XML, JSON.

The second figure (fig.3.2) is an example of a scraping code, in this case the starting URLs (start_urls variable) is extended. From each URL the text in the “p” tags will be retrieved and then yielded.

For this project is always preferred the first example (fig.3.1) as it does require less effort while achieving the same result. Unfortunately, in some cases, as we can see from the URLs in the second example, there is no common word[25] or set of words between the different links that can be used in the “allow” argument of the LinkExctrator object; in this case the programmer has to look on the company site for the URLs to be scraped and insert them manually in the bot code.

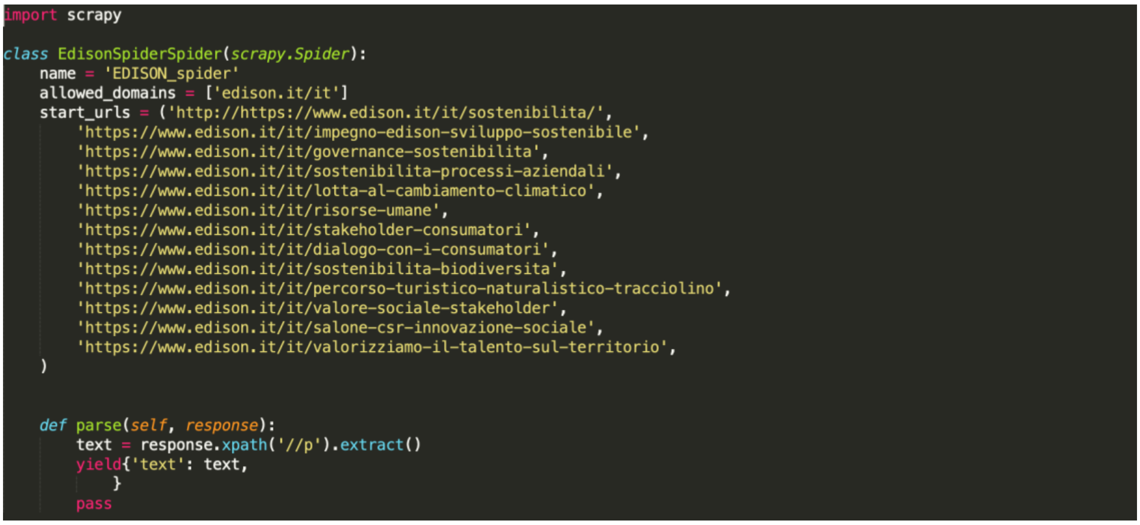


```
import scrapy
from scrapy import Spider
from scrapy.http import Request
from scrapy.spiders import CrawlSpider, Rule
from scrapy.linkextractors import LinkExtractor


class CampariSpiderSpider(CrawlSpider):
    name = 'CAMPARI_spider'
    allowed_domains = ['camparigroup.com']
    start_urls = ['http://camparigroup.com/it/page/sostenibilita/']

    rules = (Rule(LinkExtractor(allow=('sostenibilita')), callback='parse_text', follow = True),)

    def parse_text(self, response):
        text = response.xpath('//p').extract()
        yield{'text': text,
            }
        pass
```

Fig. 3.1 Example of a spider code to scrape and crawl starting from a URL.

[24] HtmlResponse and XlmResponse are classes of subclasses of TextResponse that encode support in dealing with websites by exploring the meta http-equiv attribute that reveals the information needed by the scraping bot. The TextResponse adds encoding capabilities to the Response class that can only deal with binary data, i.e. media files such as images and sound files.

[25] A high number of websites did not have a “sostenibilita” page on their site but the information were under different named pages as “/our_impact” or “/responsibility” and many more.



```
import scrapy

class EdisonSpiderSpider(scrapy.Spider):
    name = 'EDISON_spider'
    allowed_domains = ['edison.it/it']
    start_urls = ('http://https://www.edison.it/it/sostenibilita/',
        'https://www.edison.it/it/impegno-edison-sviluppo-sostenibile',
        'https://www.edison.it/it/governance-sostenibilita',
        'https://www.edison.it/it/sostenibilita-processi-aziendali',
        'https://www.edison.it/it/lotta-al-cambiamento-climatico',
        'https://www.edison.it/it/risorse-umane',
        'https://www.edison.it/it/stakeholder-consumatori',
        'https://www.edison.it/it/dialogo-con-i-consumatori',
        'https://www.edison.it/it/sostenibilita-biodiversita',
        'https://www.edison.it/it/percorso-turistico-naturalistico-tracciolino',
        'https://www.edison.it/it/valore-sociale-stakeholder',
        'https://www.edison.it/it/salone-csr-innovazione-sociale',
        'https://www.edison.it/it/valorizziamo-il-talento-sul-territorio',
    )


    def parse(self, response):
        text = response.xpath('//p').extract()
        yield{'text': text,
            }
        pass
```

Fig.3.2 Example of a spider code to scrape specific URLs

In the following section are presented the characteristics of the companies' websites and the specifics of the project creation. The code for each company is in Appendix A.

3.3 Companies' website specifics

A2A S.p.a

The home page of the company is https://www.a2a.eu/it. The company has a sustainability menu page, that is followed by other subsequential sustainability pages. I wrote the code starting from the http://a2a.eu/sostenibilita page and I allowed following pages to be retrieved if they have "sostenibilita" in their URL. After each one is crawled the text is retrieved. Reports and external links related to different sites with different domain, other than the one specified by the code, are not retrieved. The text is then stored is a company-specific csv file that then will be analyzed.

ACEA S.p.a

The company home page is https://www.gruppo.acea.it/. The sustainability page is under the name "il nostro impegno" in the menu tab. The following 11 pages in the sub menu are all retrieved, the text is extracted and stored in a csv file.

ACSM-AGAM S.p.a

The home page of the company is http://www.acsm-agam.it/home.

The company website does not present much text and only contains .pdf files that have the information related to the page searched topics. The company do not present any general sustainability page, but only a "sustainability reports" page with the 2018 and 2019 reports in pdf files. Since I didn't take any sustainability report in consideration given also the scope of the thesis, I decided to retrieve just the text from the /profilo page. In order to do it I did not use a crawler that retrieves successive webpages but only the selected page has been scraped.

Aereoporto di Bologna S.p.a

The home page of the company is https://www.bologna-airport.it. The company presents its sustainability's pages under the menu page "ambiente-qualita-e-sicurezza" where it stores all the data and comments. The code is developed to retrieve all the text from these URLs. The method used is by using a web crawler that reiterates the task of extracting the text from all the web pages. The retrieved text is stored into a csv file.

Alerion S.p.a

The home page of the company is http://www.alerion.it/home/. The company is a wind energy company. On their site there is no sustainability page but there is one that can be assumed as a proxy called "wind" where they explain their philosophy. The bot retrieves the page and collects the text from it. The site seems very dry of actual text and there is abundance of pdf links. For the purpose of the thesis I stopped at the http://www.alerion.it/wind/ page.

Amplifon S.p.a

The company investor relations page is https://corporate.amplifon.com/it. The company has a sustainability page under the "sostenibilita" word. The pages related to that word are also retrieved. The company has restrictions on the use of bots so, as for other sites, the settings are changed to retrieve the needed data.

Ascopiave S.p.a

The home page of the company is https://www.gruppoascopiave.it/. In the top line menu there is the section related to the sustainability page that contains the beliefs of the company on the matter. Other than the text there are links to the sustainability reports but, as said before, the analysis is focused on the websites and not on the reports that will not be retrieved. Given the shortness of the text, the development of a code would have been inefficient, therefore, I decided to manually copy the text without using a web scraping bot.

Atlantia S.p.a

The home page of the company is https://www.atlantia.it/. The company has a sustainability page in the menu that is then related to other three pages on the website. They all share “sostenibilita” in the URLs and given this characteristic I decided to write the crawl on this future. I retrieved the text and stored it in a csv file.

Campari S.p.a

The home page of the company is https://www.camparigroup.com/it. The company presents its sustainability page on the menu under the “sostenibilità” voice. I wrote the code to retrieve the data in the page and the following ones on the sub menu. The words are then stored in a csv file.

CNH Industrial S.p.a

The home page of the company is https://www.cnhindustrial.com/it-it/Pages/homepage.aspx. The company has a sustainability page in the menu and other related pages that have in the URL the “sustainability” word. The text is retrieved with a crawl bot and stored in a csv file for future analysis. Some settings are changed to use the bot on the site, in particular the ROBOTSTXT_OBEY is set to “False”.

DiaSorin S.p.a

The home page of the company is https://diasoringroup.com/it. On the main menu the company has the sustainability page, that is then divided in another four pages in which other “sostenibilita” URLs are present. The code of the bot retrieves all these pages. Some of the text was present, under a different tag name, in expandable parts of the visualized pages. The correct specification has been stored in a new variable of the code to retrieve it. The retrieved text is then stored into a csv file.

Edison S.p.a

The home page of the company is https://www.edison.it/it/. In the menu there is a sustainability page, that is not ideal since it mostly present other redirecting links. After an inspection of the “impegno-edison-sviluppo-sostenibile”, “governance-sostenibilita”, “sostenibilita-processi-aziendali”, “lotta-al-cambiamento-climatico”, “risorse-umane”, “stakeholder-consumatori”, “dialogo-con-i-consumatori”, “sostenibilita-biodiversita”, “percorso-turistico-naturalistico-tracciolino”, “valore-sociale-stakeholder”, “salone-csr-innovazione-sociale”, “valorizziamo-il-talento-sul-territorio” links, a code on these different pages is created to retrieve their text. The method used is not a web crawler since it would create an enormous amount of data that would not add much information compared to the noise. Given this environment all the scraped URLs are set.

Elettra Investimenti S.p.a

The home page of the company is https://www.elettrainvestimenti.it/. The company has in the top menu the sustainability page but there is near no text in it. I decided to write the code to retrieve and store it.

Enel S.p.a

The site presents its sustainability page on the investor relation page (https://www.enel.com/it/investitori), so the information will not be of easy access to clients that doesn't search for these specific qualities of the company. While the company addresses very briefly the sustainability theme on the https://www.enel.it page, it is not as near exhaustive as the content found in the former website. In this case the code starts scraping from https://www.enel.com/it/investitori/sostenibilita and retrieve all the text from the "p" tag.

ERG S.p.a

The home page of the company is https://www.erg.eu/it/home where is present the "sostenibilita" page. The bot has the allowed domain "erg.eu". The site presents a specific class, "text-content", under which the text is written, so instead of completing the retrieving process of all the text from a tag, a class is chosen. Moreover, in most ERG webpages there is the option to expand the text. In some cases the expanded text has three/four time the number of words of the not-expanded pages, so I decided to also retrieve this part of the text for every page that contain the expansion option. In particular the "expansion" class on the ERG website is "accordion-item-text".

Falckrenewables S.p.a

The home page of the company is https://www.falckrenewables.com/. It presents a menu and its sustainability page is under the name "comunità". While the page at a first look does not present any significant text, the expandable parts of the page are full of explanations on the company view on sustainability. The code is created to retrieve the shown text as well as its hidden parts.

Ferrari S.p.a

The home page of the company is https://corporate.ferrari.com/it. The sustainability page is not present in the main menu but it is under the “who we are” page. I decided to scrape the https://corporate.ferrari.com/it/chi-siamo/sostenibilita page and all the pages connected to that page that has “sostenibilita” in the URL, since there seemed to not be present other related pages with different names. From the result the only page scraped was the starting one, that suggests that no other sustainability page was present.

Frendy Energy S.p.a

The company is controlled by Edison S.p.a. The home page of the company is https://frendyenergy.edison.it/. The company present no sustainability page and a very few text parts are present on the overall site. The code retrieves the text from the front page that was the richest of information.

Generali S.p.a

The home page of the company is https://www.generali.it. The sustainability page as for Ferrari is in under the “who we are” menu voice. I wrote the code starting from the https://www.generali.it/chi-siamo/sostenibilita page and I allowed following pages to be retrieved if they have “sostenibilita” in their URL. The text from the inspection of the site is under “p” tags and “li” tags. Unfortunately, the “li” tags contain lot of strings, other than text, related to the specifications of how the site is viewed by a user. The analysis of the text will filter those non-Italian words out. The bot retrieves and store the text in a csv file.

HERA S.p.a

The company home page is https://www.gruppohera.it/. The sustainability page is present under the https://www.gruppohera.it/gruppo/sostenibilita link. The text is retrieved and stored in a csv file.

Iniziative Bresciane S.p.a

The home page of the company is http://www.iniziativebrescianespa.it/. The company does not present any sustainability page but there is a “quality and environment” page that asses how the company addresses the impact on the environment. The code retrieves all the text under the class "MsoNormal".

Intesa Sanpaolo S.p.a

The sustainability page is on the main menu from the home page of the company (https://group.intesasanpaolo.com/it/), and it is https://group.intesasanpaolo.com/it/sostenibilita. A crawling code retrieves the page and all the others in the sub menu related to it. Since there were limitation to the use of bots on the site, changes to the settings.py of the program are needed to complete the task and be as friendly as possible to the site.

IREN S.p.a

The home page of the company is https://www.gruppoiren.it/home. In the menu there is a sustainability page, but the following pages related to the topic are not followed from the “sostenibilita” URL. I then wrote the allowed domains and in particular the following ones: 'sostenibilita', 'governance-della-sostenibilita', 'strumenti-di-csr', 'matrice-di-materialita', 'contributo-agli-sdgs', 'servizi-sostenibili'. Then it is retrieved all the text under the “p” tag.

Italgas S.p.a

The home page of the company is https://www.italgas.it/it/. The company presents sustainability’s pages under the menu page “our efforts”. The code is developed to retrieve all the text from the pages related to “il-nostro-impegno”. The method used is by using a web crawler that reiterate the task of extracting the text for all the pages. Then, as all the other bots for the sites analyzed, the text is retrieved and stored into a csv file.

Moncler S.p.a

The investor relator home page is https://www.monclergroup.com/it/. The company has a sustainability page in the menu on the main page. It is then connected to 42 other pages. They all share the “sostenibilita” word in the URLs so I decided to write a bot that is able to crawl the Moncler pages containing the specific word. Given the restrictions imposed by the site I needed to modify the settings before proceeding.

NEXI S.p.a

The home page of the company is https://www.nexi.it/. The company has a sustainability page under the “who we are” page on the menu page. I decided to retrieve the main “sostenibilita” page and all the related ones that share the same word in the URLs. Given the restriction of the site in the robots.txt page of NEXI site I had to change the settings of the spider and set a request delay of 1.69 seconds.

Poste Italiane S.p.a

The home page of the company is https://www.posteitaliane.it/. The company has a sustainability page on the main menu. Under the sustainability menu there is another sub menu with topics related to E, S, G; each of which has from two to seven other voices. Given the structure of the site I could not crawl the pages that contain a common word to each page since the relative URL is different for each page and has no common words except for the base (https://www.posteitaliane.it/). I copied all the URLs connected to the sustainability topic and then retrieved all the text from them with a bot.
Since the text was under different tags and class, I created 4 categories that allows to parse the text from different locations. The scraped pages are in total 49.

Renergetica S.p.a

The home page of the company is https://www.renergetica.com/. The company produces the electricity from solar panels, there is no actual sustainability page and most of the description of the company process is in the "reti-ibride" page. I decided to retrieve it and extract the text from it. The overall site does not present a meaningful amount of text.

SNAM S.p.a

The home page of the company is https://www.snam.it/it/index.html. In the menu there is a sustainability page that is not ideal since it mostly presents other redirecting links. After an inspection of the "strategia_per_futuro/", "impegni_snam/", "responsabilita_verso_tutti/", "agire_per_ambiente/", "lavorare_in_sicurezza/", "crescere_con_fornitori/", "valorizzare_le_persone/", "reporting_e_performance/" links, the code is developed on these different pages to retrieve their text. The method used is not a web crawler since it would create an enormous amount of data that would not add much information compared to the noise created. Given this environment a spider is created.

Stellantis S.p.a

The home page of the company is https://www.stellantis.com/it. The company has a sustainability page but it presents just an overview. In the page there are also links to sustainability reports but as for the other companies these are avoided. Given the shortness of the text I manually copied the text instead of creating a code to retrieve it.

Tim S.p.a

The home page of the company is https://www.gruppotim.it/it.html. The company has a sustainability page in the menu on the main page. All the pages with the "sostenibilita" in the URLs are retrieved. In total 62 pages are crawled and the text is stored, as for all the other websites analyzed, in a csv file.

Unicredit S.p.a

The home page of the company is https://www.unicreditgroup.eu/it.html. The company has a sustainability menu page, that is followed by other pages. While the sustainability home page has a certain structure in its URL, the following pages are structured in a more friendly matter for the action of a web crawler. In particular the crawler follows the "a-sustainable-bank" common text in the URLs to retrieve all the pages related to the sustainability topics. The text under the "p" tag is retrieved and stored in a csv file.

3.4 ESG Dictionary creation

In this section are explained the main characteristics of the dictionary and its creation process in order to choose the best set of words to be compared with the individual company text retrieved by the bots.

3.4.1 The Italian language problem

The Italian language has apostrophes and accents in words such as "sostenibilità", "virtù", "legalità" and this raises a problem: the words could not be counted since accent letters are modified in the extrapolation method and during the creation of the csv file. For example, word such as "integrità", in the csv file are displayed as "integrit√†", "virtù" is "virt√π", "l'azienda è" becomes "l,Äôazienda √®".

When the analyzing code would have searched for matching words it would have missed these words into the count. In order to deal with this problem, I decided to create the dictionary using only the roots of the words.

For every accented word I dropped the miswritten letter in question and so, by continuing the previous example, I searched for "integrit" and "virt".

Also, every other word that shares the root with my "modified" words were not double counted in the dictionary in order to not incur in double counting errors. So, if in the retrieved company text appears “virtù” and “virtuosismo” by searching for “virt” I include both in the counting process without the need to add “virtuosismo” as a dictionary word.

3.4.2 ESG dictionary

By looking at different sources [94] [95] [96] [97] [98], I created the following dictionary of word roots divided into the E, S, G topics:

Topic	Dictionary
Environmental	biodiversi, biodegrad, carboni, clim, monossid, deforest, desertif, siccit, terremot, energi, alluvion, mar, fium, risors, natura, riscaldament, serra, rinnov, ozono, inquin, vulcan, spazzatur, ambient, atmosfer, eco, scart, montagn, vent, sol, acqu, suolo, verd, minor, gener, territori
Social	cambiament, monous, salvaguard, estin, sprec, ricil, atten, scrat, combatt, conserv, prote, evol, permess, preven, cambi, riform, sicur, ripristin, benefic, salut, avanza, verific, soste, proced, iniziativ, procedur, ottimiz, rinnov, sicur, standard, traspare, collabor, condivi, impegn, positiv, innova, qualit, inclus
Governance	accord, etic, diritt, dover, socioeconomic, civil, social, comunit, unit, riform, istituzion, organizzazion, leader, soci, norma, equi, comuni, accord, contratt, comitat, responsabil, maternit, paternit, management, consigli, amministr, iniziativ, cultur, person

3.5 Analyzing ESG words

In this research the analysis focuses only on single words since analyzing both, single words and group of words, would be hazardous given the high risk of double counting and the rise of tokenization[26] problems.

The risk also increases as the topic analyzed become larger and not clearly bounded as the one I'm going to analyze. In particular with the tokenization, having a broad topic (ESG) could lead to errors in the grouping words process, given that wrong groups could lead to erroneous evaluation of the text.

To decide which words to include in a vocabulary without knowing the subject the most used methodology [99] [100] in text mining is to divide the text by topic and analyze the whole text. In particular the creation of a document-term matrix [101] that is usually used to evaluate the frequencies of the terms (columns of the matrix) with respect to the documents analyzed (rows of the matrix), can also be easily applied to segments of texts. The vector space model is used to create the document-term matrix with which is analyzed the frequency and the importance of the single term for the context analyzed and the method used. This approach was proposed by Salton et al. [102] that stated that the value of a system is expressed by a function of the density of the object space. Their model is mostly known as the term frequency-inverse document frequency model: the term frequency, tf $(n, d) = f_{n,d}$, that will also be used in this work, is obtained by counting the number of n -words that occur in the specific text (d), retrieved from the company websites, and adjust them by the length of the texts (l).

The inverse document frequency $(idf(n, d))$ is the evaluation of how much a word describe the topic analyzed, this is done by calculating the logarithm of the division between the total number of segments and the number of segments that contain the specific term, formally:

$$idf(n, d) = \log(\frac{D}{d_n}) = -\log(\frac{d_n}{D})$$

D is the total number of documents/segments, while d_t is the total amount of documents/segments that contain the n-term recurrences.

[26] A token is a sequence of words separated by a delimiter, such as spaces, points or commas. Algorithms, such as the Sentence Boundary Detection algorithms, have to be used in order to create a group of words that are coherent with the topic analyzed.

There are also more complex methods of topic models as LSI or LSA, that use a technique called singular value decomposition (SVD) to establish patterns between words but are out of the scope of this work. The inverse document frequency won’t be used here since the topic is already defined by the web scraping code and an idf inclusion would distort the data.

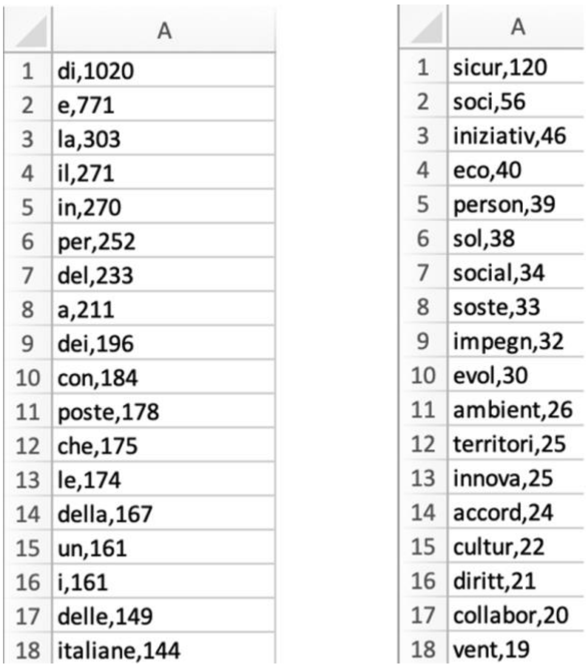
In the path chosen to analyze the words in the websites, a signal to noise tradeoff [103] has been taken into account since the ESG topics have a high risk of double counting, given also the shaded lines between the topics and the higher noise related to deep parsing.

3.5.1 Data cleaning and word frequency

The drawback of shortening the time required for coding hardcoded spiders that would have been able to retrieve and store structured data, is that the .csv files on which the texts are saved have to be manipulated in order to perform an errorless counting process. It is developed a code (Appendix B) that calculates the frequencies of all the words for each text file starting from a folder containing all the companies’ CSV files. In particular, before beginning the counting process the retrieved text of each company has to be manipulated in order to avoid bugs. The code starts by eliminating sequences of characters that are not related to the Italian vocabulary, such as “<.*?>”, “””, “\.”. Then the spaces at the beginning and at the end of each line are eliminated, each uppercase character is made lowercase and lastly the phrases are divided into single words. Two new csv files for each company are created with the following rows structure: “word_name,word_name_frequency”.

The first one, called “name_top50”, containing up to 50’000 single words of the text extracted, is ordered by frequency, the second file that contains the dictionary words is ordered in the same method. The code simply adds +1 to the already existing word count to print the total number for each word. If a word is not present in the CSV file the code starts the count from 1 when the bot encounters it. The process is the same for the second CSV file, called “name_selected”, but only the words that contain the words’ roots in the dictionary (the “selected_words” variable of the code) and that match the scraped words of the text of the company specific CSV file are added to the file.

The following figure (fig.3.3) is a representation of how the files are created. In particular, on the left there is an example of the first CSV file, while on the right there is the “name_selected” file.



	A
1	di,1020
2	e,771
3	la,303
4	il,271
5	in,270
6	per,252
7	del,233
8	a,211
9	dei,196
10	con,184
11	poste,178
12	che,175
13	le,174
14	della,167
15	un,161
16	i,161
17	delle,149
18	italiane,144

	A
1	sicur,120
2	soci,56
3	iniziativ,46
4	eco,40
5	person,39
6	sol,38
7	social,34
8	soste,33
9	impegn,32
10	evol,30
11	ambient,26
12	territori,25
13	innova,25
14	accord,24
15	cultur,22
16	diritt,21
17	collabor,20
18	vent,19

Fig.3.3 Example of the two created CSV files: “name_top50” and “name_selected” files

3.6 Frequency plots

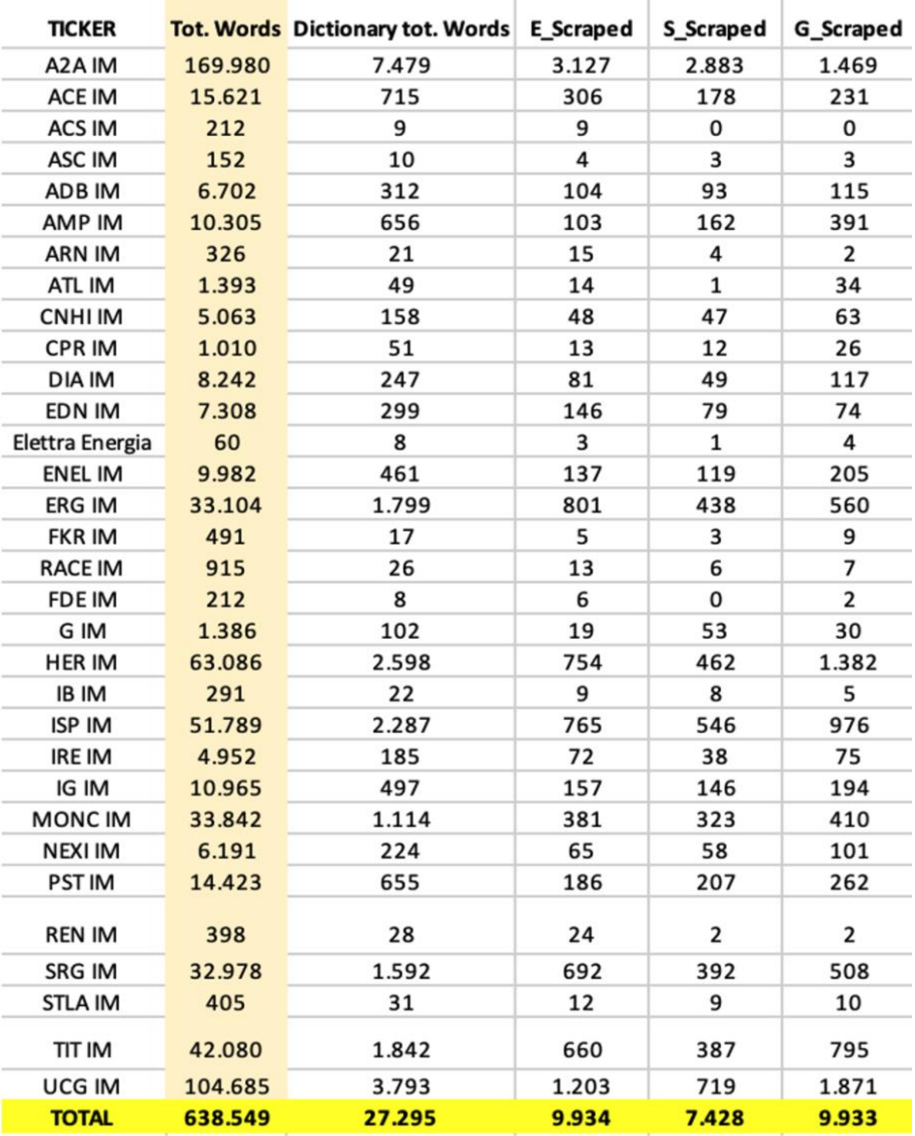
A new code (Appendix C) is created to plot on bar charts the CSV files for each company. The figures in Appendix D are the result of the execution of the code. The first image for each company is plotted by using the “name_top50” files, while the second by using the “name_selected” files. Given the unreadable figures that would result by plotting 50’000 words on the X-axes each figure contains only the 50 most frequent words.

The most common words are not contextual, as seen from the “name_top50” files’ figures. Usually stopwords’ lists, such as “and”, “the”, “it”, etc. are used to filter them out. These lists could be easily found online from different sources. Other than clearing the data, the elimination process reduces the amount of computational power required to analyze the text. This process is extremely useful when the project scope is to understand the topic of a text. In this research the topics are already known and the creation of a specific dictionary is more useful for the analysis. The bar charts of the “name_selected” files show the distributions ordered by frequency for each company.

3.7 Environmental, social and governance ratings

A new code (Appendix E) is created to evaluate the focus by topic of each company analyzed. In particular, the previously created dictionary is now divided into three groups: environmental, social and governance. Each word in the “name_selected” file is now added to one of the specific subgroups created. The different values between each topic suggest the different focuses the companies have on the specific subjects.

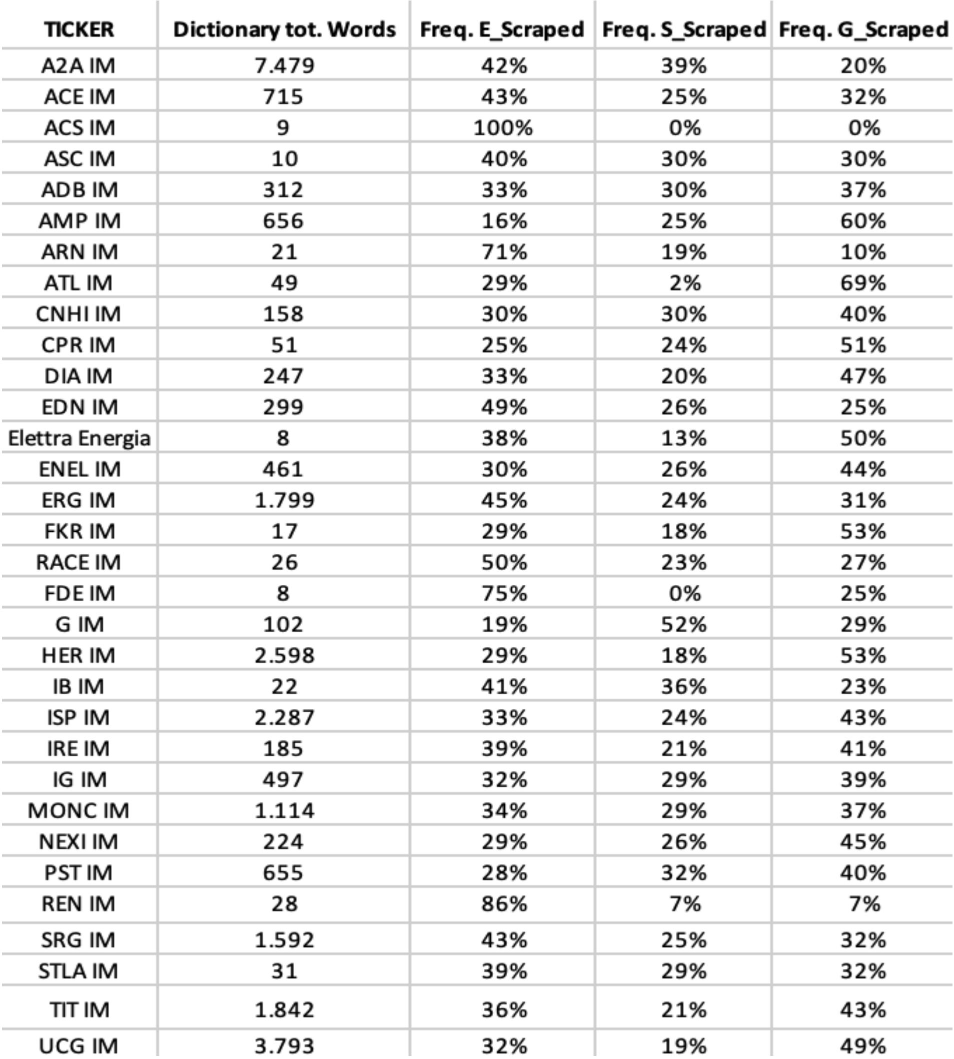
The results are presented in the figure below (Fig. 3.4). Starting from left to right, the first column contains the tickers of the analyzed companies; the second contains the total extracted words; then is listed the total number of words from the dictionary selected pool and, in the last three columns, there are the words divided by topic.



TICKER	Tot. Words	Dictionary tot. Words	E_Scraped	S_Scraped	G_Scraped
A2A IM	169.980	7.479	3.127	2.883	1.469
ACE IM	15.621	715	306	178	231
ACS IM	212	9	9	0	0
ASC IM	152	10	4	3	3
ADB IM	6.702	312	104	93	115
AMP IM	10.305	656	103	162	391
ARN IM	326	21	15	4	2
ATL IM	1.393	49	14	1	34
CNHI IM	5.063	158	48	47	63
CPR IM	1.010	51	13	12	26
DIA IM	8.242	247	81	49	117
EDN IM	7.308	299	146	79	74
Elettra Energia	60	8	3	1	4
ENEL IM	9.982	461	137	119	205
ERG IM	33.104	1.799	801	438	560
FKR IM	491	17	5	3	9
RACE IM	915	26	13	6	7
FDE IM	212	8	6	0	2
G IM	1.386	102	19	53	30
HER IM	63.086	2.598	754	462	1.382
IB IM	291	22	9	8	5
ISP IM	51.789	2.287	765	546	976
IRE IM	4.952	185	72	38	75
IG IM	10.965	497	157	146	194
MONC IM	33.842	1.114	381	323	410
NEXI IM	6.191	224	65	58	101
PST IM	14.423	655	186	207	262
REN IM	398	28	24	2	2
SRG IM	32.978	1.592	692	392	508
STLA IM	405	31	12	9	10
TIT IM	42.080	1.842	660	387	795
UCG IM	104.685	3.793	1.203	719	1.871
TOTAL	638.549	27.295	9.934	7.428	9.933

Fig.3.4 Companies' results from the web scraping process

By looking at the above results there are 9 companies that have less than 500 words retrieved by the bot, 8 companies have more than a thousand words from my selected dictionary. 16 company, of which 6 are in the utility sector, are more focused on the Governance aspect of the sustainability topic. 15 companies, of which 11 are in the utility sector, are focused on the Environmental aspect and only one company is focused more on the Social topic and it is Generali S.p.a (G.IM). The following figure (fig.3.5) presents the percentage of words, for topic, of the total dictionary words.



TICKER	Dictionary tot. Words	Freq. E_Scraped	Freq. S_Scraped	Freq. G_Scraped
A2A IM	7.479	42%	39%	20%
ACE IM	715	43%	25%	32%
ACS IM	9	100%	0%	0%
ASC IM	10	40%	30%	30%
ADB IM	312	33%	30%	37%
AMP IM	656	16%	25%	60%
ARN IM	21	71%	19%	10%
ATL IM	49	29%	2%	69%
CNHI IM	158	30%	30%	40%
CPR IM	51	25%	24%	51%
DIA IM	247	33%	20%	47%
EDN IM	299	49%	26%	25%
Elettra Energia	8	38%	13%	50%
ENEL IM	461	30%	26%	44%
ERG IM	1.799	45%	24%	31%
FKR IM	17	29%	18%	53%
RACE IM	26	50%	23%	27%
FDE IM	8	75%	0%	25%
G IM	102	19%	52%	29%
HER IM	2.598	29%	18%	53%
IB IM	22	41%	36%	23%
ISP IM	2.287	33%	24%	43%
IRE IM	185	39%	21%	41%
IG IM	497	32%	29%	39%
MONC IM	1.114	34%	29%	37%
NEXI IM	224	29%	26%	45%
PST IM	655	28%	32%	40%
REN IM	28	86%	7%	7%
SRG IM	1.592	43%	25%	32%
STLA IM	31	39%	29%	32%
TIT IM	1.842	36%	21%	43%
UCG IM	3.793	32%	19%	49%

Fig.3.5 Frequency of the E, S, G topics on the total dictionary words

3.8 Financial performance measures

In this section are presented the measures that will be compared to the bot extracted data to check if a relation between the different topics exits. In particular, the first subsection illustrates the stocks performances of the analyzed companies, while the second presents the Bloomberg extracted data.

3.8.1 Stocks returns of the analyzed companies

The following table presents the last 12 months[27] and the last 5 years[28] stock returns for the analyzed companies. The latter, "Last 5Y Returns" column, does not contain values

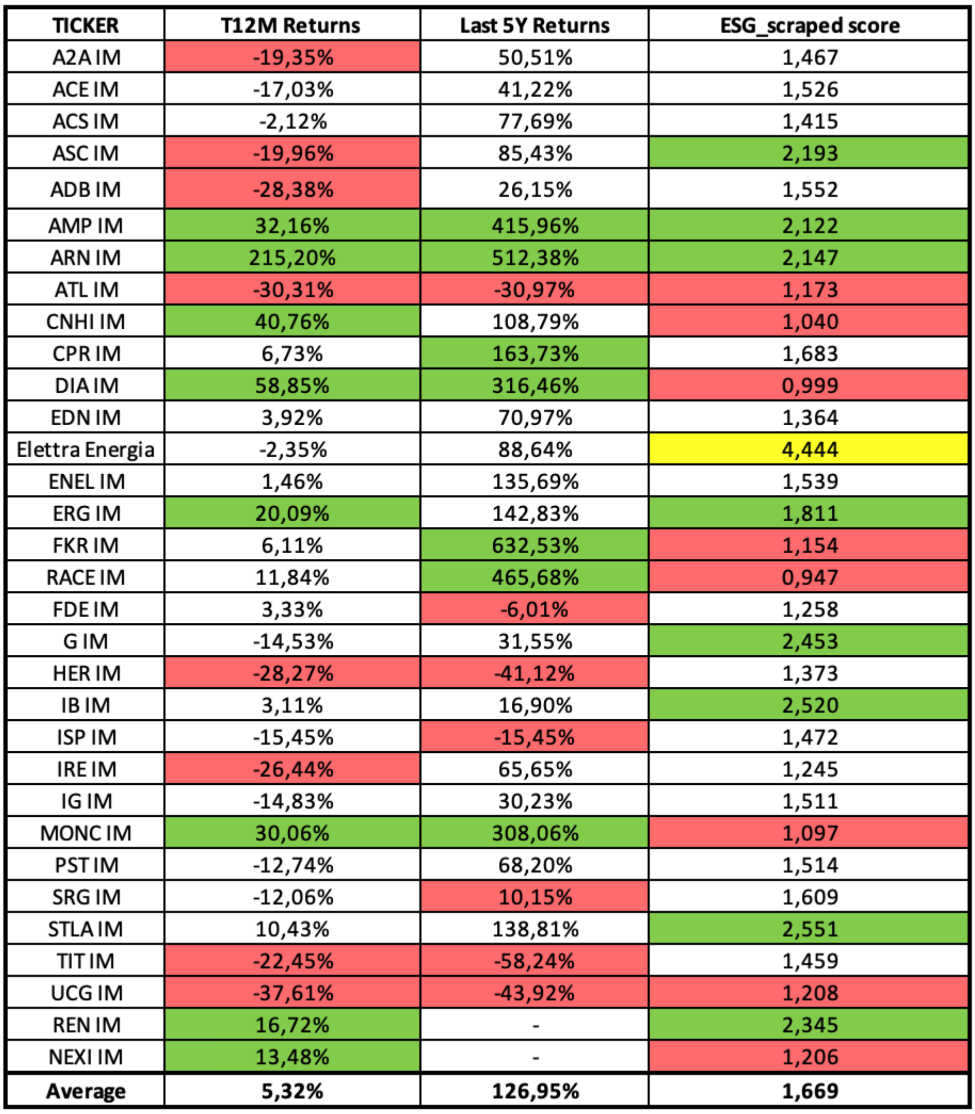
[27] From the 10th of February 2020 to the 10th of February 2021

[28] From the 10th of February 2016 to the 10th of February 2021

for Renergetica S.p. and Nexi S.p.a since the listing was on the 7th of August 2018 for the first company and on the 18th of April 2019 for the second one.

Moreover, the first and the fourth quartile for each column are highlighted in red and green respectively. The ESG scores column presents a yellow box as Elettra Investimenti S.p.a ("Elettra Energia" in the table), that would be in the fourth quartile, has the highest score but the lowest number of words scraped from its website (i.e. 60 words). By looking at the table it can be noted that half of the best returns (i.e. fourth quartile) on a five year time frame are also in the best quartile of the returns in the one year time frame. The same can be said about the worst performing stocks from the sample used. Moreover, there does not seems to be any relevant relation between the quartiles of the ESG scores and the performances on both the time frames.

Further analysis will be performed in the following sections to mathematically check if a relation can be found.



TICKER	T12M Returns	Last 5Y Returns	ESG_scraped score
A2A IM	-19,35%	50,51%	1,467
ACE IM	-17,03%	41,22%	1,526
ACS IM	-2,12%	77,69%	1,415
ASC IM	-19,96%	85,43%	2,193
ADB IM	-28,38%	26,15%	1,552
AMP IM	32,16%	415,96%	2,122
ARN IM	215,20%	512,38%	2,147
ATL IM	-30,31%	-30,97%	1,173
CNHI IM	40,76%	108,79%	1,040
CPR IM	6,73%	163,73%	1,683
DIA IM	58,85%	316,46%	0,999
EDN IM	3,92%	70,97%	1,364
Elettra Energia	-2,35%	88,64%	4,444
ENEL IM	1,46%	135,69%	1,539
ERG IM	20,09%	142,83%	1,811
FKR IM	6,11%	632,53%	1,154
RACE IM	11,84%	465,68%	0,947
FDE IM	3,33%	-6,01%	1,258
G IM	-14,53%	31,55%	2,453
HER IM	-28,27%	-41,12%	1,373
IB IM	3,11%	16,90%	2,520
ISP IM	-15,45%	-15,45%	1,472
IRE IM	-26,44%	65,65%	1,245
IG IM	-14,83%	30,23%	1,511
MONC IM	30,06%	308,06%	1,097
PST IM	-12,74%	68,20%	1,514
SRG IM	-12,06%	10,15%	1,609
STLA IM	10,43%	138,81%	2,551
TIT IM	-22,45%	-58,24%	1,459
UCG IM	-37,61%	-43,92%	1,208
REN IM	16,72%	-	2,345
NEXI IM	13,48%	-	1,206
Average	5,32%	126,95%	1,669

Table 3.1 Stocks returns, evidence of first and fourth quartiles

3.8.2 Bloomberg's ratings and financial metrics

There is evidence [104] [105] that different ratings methodologies lead to statistically different results and scores. In particular, even if the analyzed concepts are the same, the different components and their weightings do not create a converging opinion on the single analyzed companies. This disagreement is not helpful for the investors that want to choose sustainable companies to construct their portfolio. The inconsistence of the selection of stocks could lead to different utility' functions and weaken the effects of the exclusion and taste premia shown in the Zerbib's model in section 1.8 (sustainable CAPM approach).

The ESG scoring methodology and weighting of factors are proprietary of the rating companies. In this project are used the Bloomberg ESG scores that ranges from 0.1 to 100 in relation to the amount of ESG disclosed data by the companies. The maximum score is assigned if every sustainability data point can be collected by Bloomberg. The data points are weighted by importance (greenhouse gas emissions carrying the highest weight) and evaluated based on the company's industry sector. Each score is based on the company reported data and do not measure the sustainability performances year over year on any data point.

Bloomberg also discloses the most weighted factor for the individual E, S, G score; they are respectively greenhouse gas emissions, workforce and board of directors' characteristics.

Regarding the financial metrics four measures are taken into account to evaluate the relation with the ESG ratings: the Net Debt/EBITDA that is a debt ratio, used as a measurement of leverage levels; the current ratio[29], a liquidity ratio and a measurement of a company ability to repay its short-term obligations; the ROA[30] that is a profitability ratio and allows to evaluate the ability of the management to generate earnings from the usage of assets and lastly the EV / T12M[31] EBITDA, a value ratio, used to evaluate the relation of the company value and its ESG rating. The enterprise value[32] is calculated as of the 10th of February 2021.

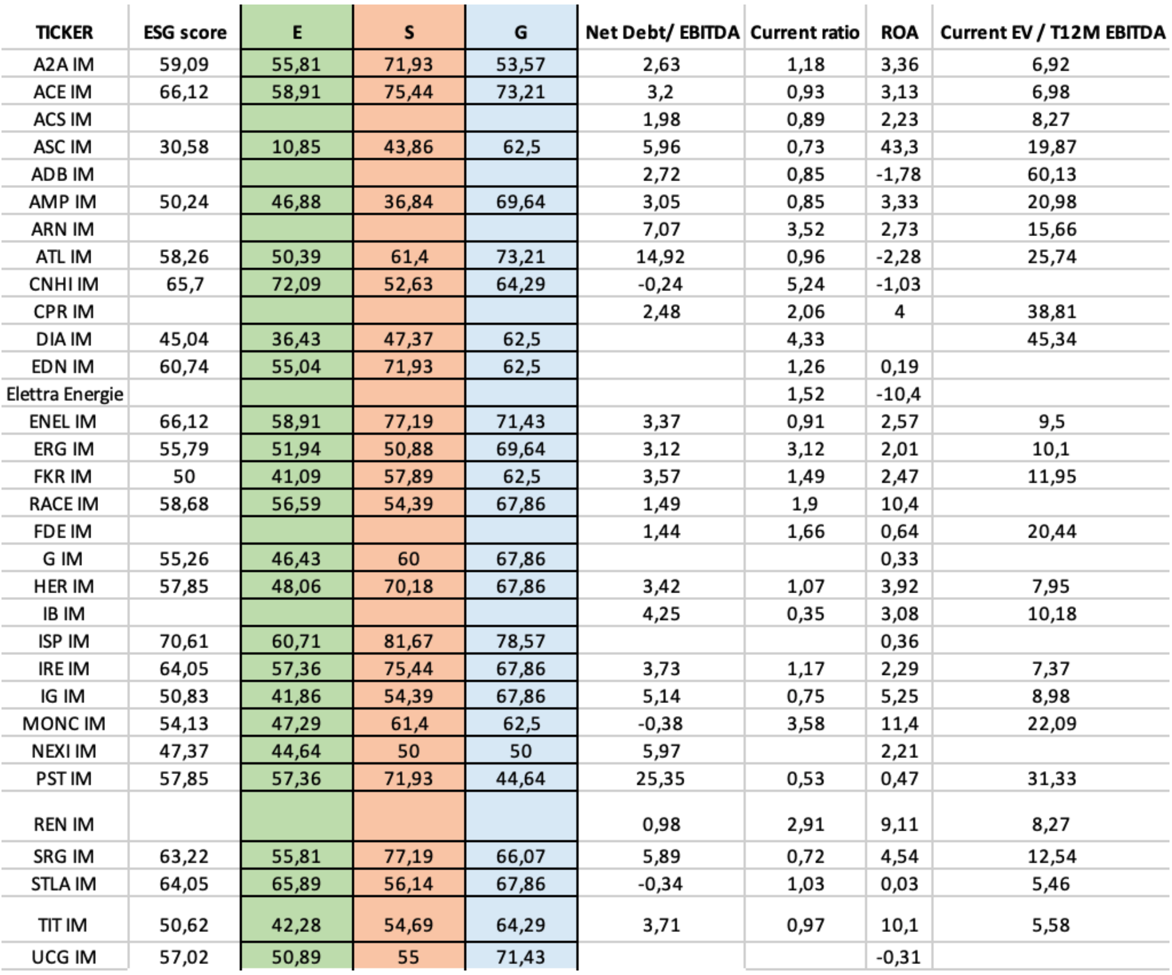
[29] Current Assets / Current Liabilities

[30] Calculated as Net Income / Total Assets

[31] T12M stands for trailing 12 months.

[32] Calculated as Market Capitalization - Cash & Cash equivalent + Preferred Equity + Minority Interest + Total Debt.

The figure below contains the values of the analyzed companies, extracted from the Bloomberg Terminal.



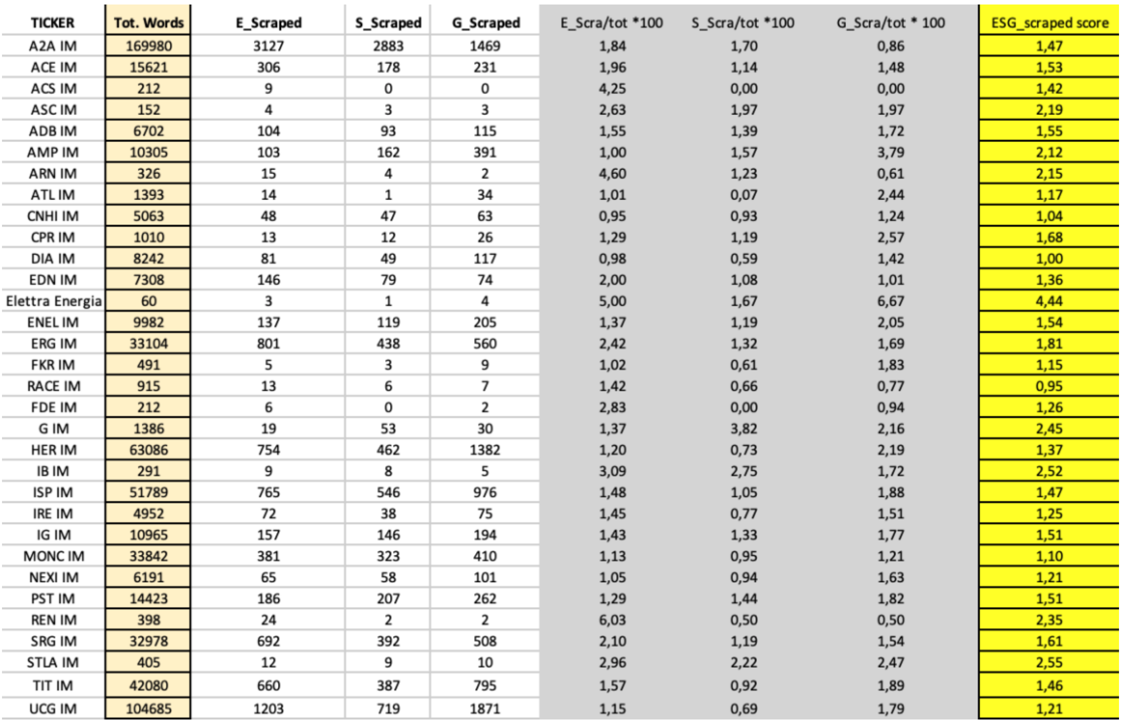
TICKER	ESG score	E	S	G	Net Debt/ EBITDA	Current ratio	ROA	Current EV / T12M EBITDA
A2A IM	59,09	55,81	71,93	53,57	2,63	1,18	3,36	6,92
ACE IM	66,12	58,91	75,44	73,21	3,2	0,93	3,13	6,98
ACS IM					1,98	0,89	2,23	8,27
ASC IM	30,58	10,85	43,86	62,5	5,96	0,73	43,3	19,87
ADB IM					2,72	0,85	-1,78	60,13
AMP IM	50,24	46,88	36,84	69,64	3,05	0,85	3,33	20,98
ARN IM					7,07	3,52	2,73	15,66
ATL IM	58,26	50,39	61,4	73,21	14,92	0,96	-2,28	25,74
CNHI IM	65,7	72,09	52,63	64,29	-0,24	5,24	-1,03	
CPR IM					2,48	2,06	4	38,81
DIA IM	45,04	36,43	47,37	62,5		4,33		45,34
EDN IM	60,74	55,04	71,93	62,5		1,26	0,19	
Elettra Energie						1,52	-10,4	
ENEL IM	66,12	58,91	77,19	71,43	3,37	0,91	2,57	9,5
ERG IM	55,79	51,94	50,88	69,64	3,12	3,12	2,01	10,1
FKR IM	50	41,09	57,89	62,5	3,57	1,49	2,47	11,95
RACE IM	58,68	56,59	54,39	67,86	1,49	1,9	10,4	
FDE IM					1,44	1,66	0,64	20,44
G IM	55,26	46,43	60	67,86			0,33	
HER IM	57,85	48,06	70,18	67,86	3,42	1,07	3,92	7,95
IB IM					4,25	0,35	3,08	10,18
ISP IM	70,61	60,71	81,67	78,57			0,36	
IRE IM	64,05	57,36	75,44	67,86	3,73	1,17	2,29	7,37
IG IM	50,83	41,86	54,39	67,86	5,14	0,75	5,25	8,98
MONC IM	54,13	47,29	61,4	62,5	-0,38	3,58	11,4	22,09
NEXI IM	47,37	44,64	50	50	5,97		2,21	
PST IM	57,85	57,36	71,93	44,64	25,35	0,53	0,47	31,33
REN IM					0,98	2,91	9,11	8,27
SRG IM	63,22	55,81	77,19	66,07	5,89	0,72	4,54	12,54
STLA IM	64,05	65,89	56,14	67,86	-0,34	1,03	0,03	5,46
TIT IM	50,62	42,28	54,69	64,29	3,71	0,97	10,1	5,58
UCG IM	57,02	50,89	55	71,43			-0,31	

Fig.3.6 Bloomberg extracted data

3.9 Cross Sectional regression analysis

As shown in the previous figure (fig.3.6) there are missing values in the data set. In the regression analysis that are performed all the companies that do not have the analyzed data for the specific regression are discarded from the computation.

It is created an ESG_scraped score from the scraped values in figure 3.7 that is calculated as the mean of the individual E_scraped, S_scraped, G_scraped scores divided by the total number of words and multiplied by 100 (the grey columns in the image below).



TICKER	Tot. Words	E_Scraped	S_Scraped	G_Scraped	E_Scra/tot *100	S_Scra/tot *100	G_Scra/tot * 100	ESG_scraped score
A2A IM	169980	3127	2883	1469	1,84	1,70	0,86	1,47
ACE IM	15621	306	178	231	1,96	1,14	1,48	1,53
ACS IM	212	9	0	0	4,25	0,00	0,00	1,42
ASC IM	152	4	3	3	2,63	1,97	1,97	2,19
ADB IM	6702	104	93	115	1,55	1,39	1,72	1,55
AMP IM	10305	103	162	391	1,00	1,57	3,79	2,12
ARN IM	326	15	4	2	4,60	1,23	0,61	2,15
ATL IM	1393	14	1	34	1,01	0,07	2,44	1,17
CNHI IM	5063	48	47	63	0,95	0,93	1,24	1,04
CPR IM	1010	13	12	26	1,29	1,19	2,57	1,68
DIA IM	8242	81	49	117	0,98	0,59	1,42	1,00
EDN IM	7308	146	79	74	2,00	1,08	1,01	1,36
Elettra Energia	60	3	1	4	5,00	1,67	6,67	4,44
ENEL IM	9982	137	119	205	1,37	1,19	2,05	1,54
ERG IM	33104	801	438	560	2,42	1,32	1,69	1,81
FKR IM	491	5	3	9	1,02	0,61	1,83	1,15
RACE IM	915	13	6	7	1,42	0,66	0,77	0,95
FDE IM	212	6	0	2	2,83	0,00	0,94	1,26
G IM	1386	19	53	30	1,37	3,82	2,16	2,45
HER IM	63086	754	462	1382	1,20	0,73	2,19	1,37
IB IM	291	9	8	5	3,09	2,75	1,72	2,52
ISP IM	51789	765	546	976	1,48	1,05	1,88	1,47
IRE IM	4952	72	38	75	1,45	0,77	1,51	1,25
IG IM	10965	157	146	194	1,43	1,33	1,77	1,51
MONC IM	33842	381	323	410	1,13	0,95	1,21	1,10
NEXI IM	6191	65	58	101	1,05	0,94	1,63	1,21
PST IM	14423	186	207	262	1,29	1,44	1,82	1,51
REN IM	398	24	2	2	6,03	0,50	0,50	2,35
SRG IM	32978	692	392	508	2,10	1,19	1,54	1,61
STLA IM	405	12	9	10	2,96	2,22	2,47	2,55
TIT IM	42080	660	387	795	1,57	0,92	1,89	1,46
UCG IM	104685	1203	719	1871	1,15	0,69	1,79	1,21

Fig.3.7 ESG score from E, S, G scraped words

Fourteen regressions are performed in order to test the hypothesis of a relation between web scraped metrics and the actual companies' data.

In particular, it is performed an exploratory analysis with the scatter plots (Appendix F) of the variables analyzed for each regression. From the analysis there seems to be no evidence of a linear relation.

In particular, it is analyzed the relation between Bloomberg scores and the scores resulting from the scraping process. The results in table 3.2 shows that the ESG_scraped scores are not a good predictor of the Bloomberg ESG scores. This is also confirmed by the coefficient of determination R^2 of 0,3 and 0,2 of the first two regression, meaning that only 3% and 2% of the variance of the ESG scores can be explained by the set of E, S, G scraped scores and the overall ESG scraped score respectively. Therefore, it is concluded that there is no linear relationship between the analyzed variables.

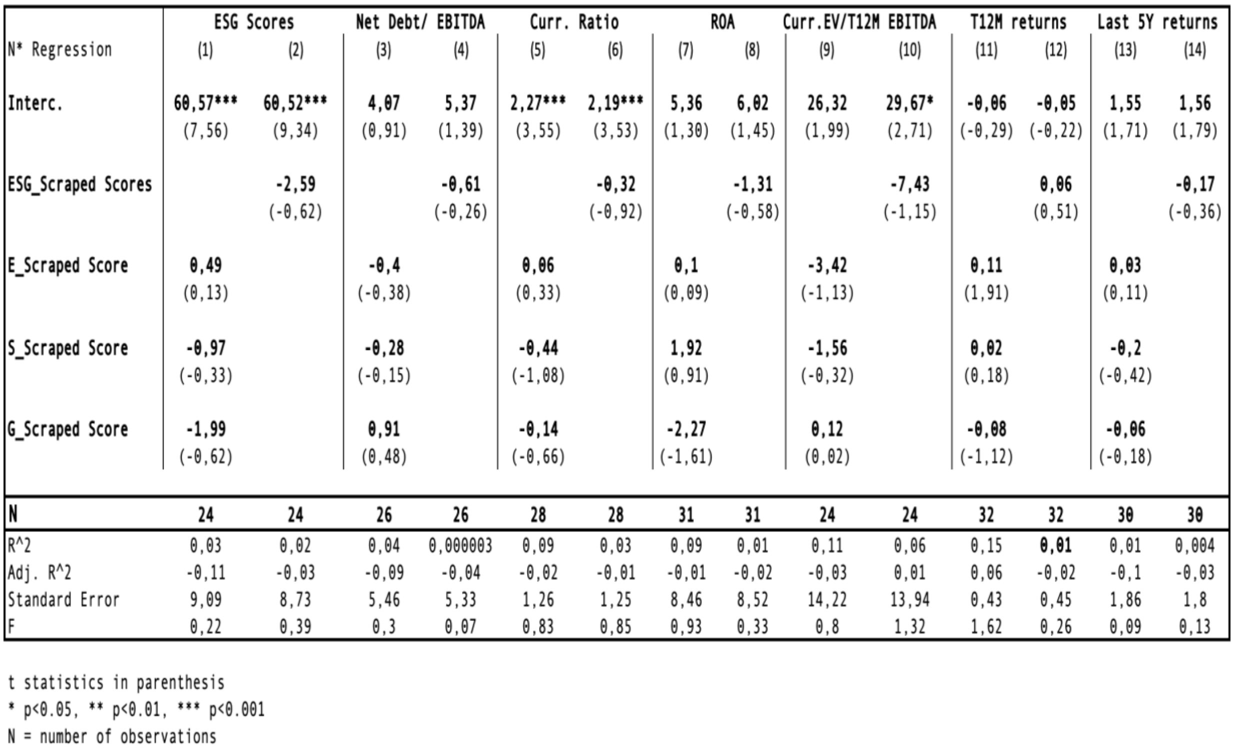
Then it is evaluated if relations between a debt ratio (Net Debt/ EBITDA), a liquidity ratio (current ratio) and a profitability ratio (ROA) and the scores resulting from the scraping process exist. The ESG_scraped scores and the single E, S, G scraped scores are not good predictors of the debt level, the liquidity levels and profitability levels of the analyzed companies.

Moreover, there is no relation between the levels of sustainability calculated with the scraping method and the values of a companies calculated as Current EV / T12M EBITDA as of the 10th of February 2021.

Lastly, four regression analysis are computed to evaluate the relation between the ESG scraped scores and the companies’ stock returns, calculated on a one and five-year time frame. It is concluded that there are not linear relations.

It is also worth noting that, given the chosen one-year time frame, there is no linear relation between the scraped sustainability levels and the stock returns after the Covid-19 shock in March 2020.

Below it is presented the table that summarize the regressions’ results. Further data can be found on the Appendix F.



	ESG Scores		Net Debt/ EBITDA		Curr. Ratio		ROA		Curr.EV/T12M EBITDA		T12M returns		Last 5Y returns	
N* Regression	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Interc.	**60,57*** **	**60,52*** **	**4,07**	**5,37**	**2,27*** **	**2,19*** **	**5,36**	**6,02**	**26,32**	**29,67***	**-0,06**	**-0,05**	**1,55**	**1,56**
	(7,56)	(9,34)	(0,91)	(1,39)	(3,55)	(3,53)	(1,30)	(1,45)	(1,99)	(2,71)	(-0,29)	(-0,22)	(1,71)	(1,79)
ESG_Scraped Scores		**-2,59**		**-0,61**		**-0,32**		**-1,31**		**-7,43**		**0,06**		**-0,17**
		(-0,62)		(-0,26)		(-0,92)		(-0,58)		(-1,15)		(0,51)		(-0,36)
E_Scraped Score	**0,49**		**-0,4**		**0,06**		**0,1**		**-3,42**		**0,11**		**0,03**	
	(0,13)		(-0,38)		(0,33)		(0,09)		(-1,13)		(1,91)		(0,11)	
S_Scraped Score	**-0,97**		**-0,28**		**-0,44**		**1,92**		**-1,56**		**0,02**		**-0,2**	
	(-0,33)		(-0,15)		(-1,08)		(0,91)		(-0,32)		(0,18)		(-0,42)	
G_Scraped Score	**-1,99**		**0,91**		**-0,14**		**-2,27**		**0,12**		**-0,08**		**-0,06**	
	(-0,62)		(0,48)		(-0,66)		(-1,61)		(0,02)		(-1,12)		(-0,18)	
N	**24**	**24**	**26**	**26**	**28**	**28**	**31**	**31**	**24**	**24**	**32**	**32**	**30**	**30**
R^2	0,03	0,02	0,04	0,000003	0,09	0,03	0,09	0,01	0,11	0,06	0,15	**0,01**	0,01	0,004
Adj. R^2	-0,11	-0,03	-0,09	-0,04	-0,02	-0,01	-0,01	-0,02	-0,03	0,01	0,06	-0,02	-0,1	-0,03
Standard Error	9,09	8,73	5,46	5,33	1,26	1,25	8,46	8,52	14,22	13,94	0,43	0,45	1,86	1,8
F	0,22	0,39	0,3	0,07	0,83	0,85	0,93	0,33	0,8	1,32	1,62	0,26	0,09	0,13

t statistics in parenthesis
* p<0.05, ** p<0.01, *** p<0.001
N = number of observations

Table 3.2 Summary of linear regressions results

3.10 Rank correlation of ESG scores on ESG_scraped scores

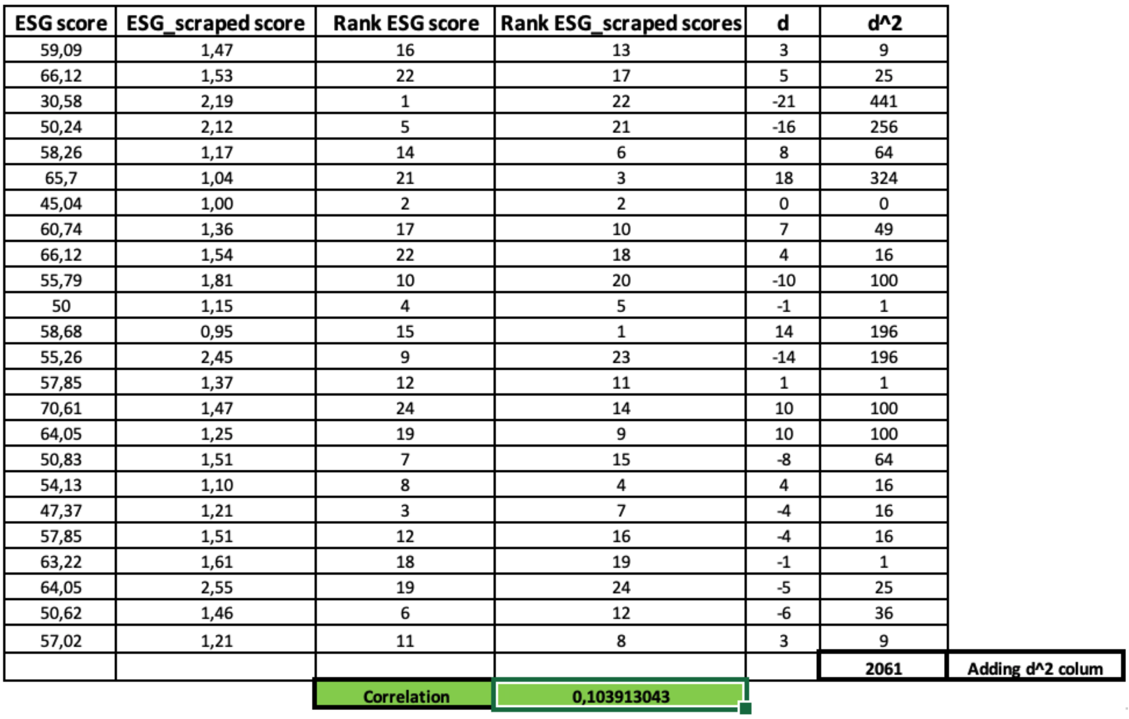
To check if the relation between the Bloomberg scores and the scraped scores is nonlinear it is calculated the Spearman’s rank correlation coefficient. It is a nonparametric measure of rank correlation, meaning that the sampling on which the computation is based does not require any specific joint probability distribution assumptions of the analyzed series. Moreover, a perfect correlation is found if the variables are related by a monotonic function. The Pearson correlation on the other hand gives a perfect correlation value if the function is linear. The Spearman correlation coefficient ranges from -1 to +1, indicating respectively a perfect negative and a perfect positive association of ranks. 0

indicates no association between the ranks. In order to perform the calculation, the analyzed series have to be ordered by rank.

By looking at the scatter plot of the first figure in Appendix F there seems that there could be a monotonic relation between the ESG scores and the ESG_scraped scores. In order to evaluate this relation the Spearman's rank correlation coefficient is calculated. In fig.3.8 each series is ordered by lowest to highest value in the "Rank" columns. Then is calculated the difference in column "d" and it is squared in the last column. The result of all the added values is 2061. The correlation is now calculated by using the following formula:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)},$$

n is the sample size analyzed, in this case it is 24. The Spearman's correlation is 0,10 and therefore there is a small, positive effect between these two variables.



ESG score	ESG_scraped score	Rank ESG score	Rank ESG_scraped scores	d	d^2	
59,09	1,47	16	13	3	9	
66,12	1,53	22	17	5	25	
30,58	2,19	1	22	-21	441	
50,24	2,12	5	21	-16	256	
58,26	1,17	14	6	8	64	
65,7	1,04	21	3	18	324	
45,04	1,00	2	2	0	0	
60,74	1,36	17	10	7	49	
66,12	1,54	22	18	4	16	
55,79	1,81	10	20	-10	100	
50	1,15	4	5	-1	1	
58,68	0,95	15	1	14	196	
55,26	2,45	9	23	-14	196	
57,85	1,37	12	11	1	1	
70,61	1,47	24	14	10	100	
64,05	1,25	19	9	10	100	
50,83	1,51	7	15	-8	64	
54,13	1,10	8	4	4	16	
47,37	1,21	3	7	-4	16	
57,85	1,51	12	16	-4	16	
63,22	1,61	18	19	-1	1	
64,05	2,55	19	24	-5	25	
50,62	1,46	6	12	-6	36	
57,02	1,21	11	8	3	9	
					2061	Adding d^2 colum
		Correlation	0,103913043			

Fig.3.8 Rank correlation variables and Spearman's correlation

Conclusions

The performed analysis evaluated more than 600’000 words from a total of 32 Italian companies. The results show that, based on the analyzed sample, there is no linear and rank correlation between how a company decides to market itself on its website and the actual E, S, G scores. Moreover, it is not found any significant relation between the ESG scores from web scraping and the debt, liquidity and profitability levels of the analyzed companies. Also, there is no linear relation between the value of the companies and their ESG scraped scores. Lastly, there is no linear relation between the scraped sustainability scores and the stock returns in a one year and a five-year time frame.

To my knowledge, prior to this project, there has not been conducted any research on the relation between the website text used by companies and their actual level of sustainability. Given the obtained results ESG investors should have particular attention to distinguish the perception of sustainability that a company wants to externalize and the actual level of it. The instruments created in this thesis could be useful to companies and investment funds to evaluate different topics on targeted websites by changing the sustainability dictionary in the code section.

Further text analysis could take into account the companies’ sustainability reports; while this seems trivial, evaluation of numbers in an unstructured text with a fully automated process could be challenging. Moreover, the analysis performed on the Italian companies could be protracted to companies operating in other countries.

Appendix A

Web Scraping Python code

It is presented the web scraping code to retrieve all the text words from the different companies' websites. The code related to Items.py, Pipelines.py and Middlewares.py is unchanged for all the companies analyzed. They are standardized by Scrapy and in my process there has been no need to change them. I will write them at the beginning of this Appendix. When using the code it has to be substituted " *NameOfTheProject* " with the actual name given to the project.

The Settings.py code has been mostly unchanged from the default one but when needed it has been modified the User_Agent, the Download_Delay and the Robotstxt_obey rules. The first one was changed if the agent of the bot was banned from the site analyzed, the delay in the download and the robots' rules (set as True or False) were imposed to be more site-friendly and to not trigger restricting mechanism that would have affected the bot performances. Then it is written the spider code for every analyzed company.

Items.py

```
import scrapy
class *NameOfTheProject*Item(scrapy.Item):
    # define the fields for your item here like:
    # name = scrapy.Field()
    pass
```

Pipelines.py

```
from itemadapter import ItemAdapter
class *NameOfTheProject*Pipeline:
    def process_item(self, item, spider):
        return item
```

Middlewares.py

```
from scrapy import signals
# useful for handling different item types with a single interface
from itemadapter import is_item, ItemAdapter
class *NameOfTheProject*SpiderMiddleware:
    # Not all methods need to be defined. If a method is not defined, scrapy
    #acts as if the spider middleware does not modify the passed objects.
    @classmethod
    def from_crawler(cls, crawler):
        # This method is used by Scrapy to create your spiders.
        s = cls()
        crawler.signals.connect(s.spider_opened, signal=signals.spider_opened)
        return s
    def process_spider_input(self, response, spider):
       # Called for each response that goes through the spider
       # middleware and into the spider.
       # Should return None or raise an exception.
        return None
    def process_spider_output(self, response, result, spider):
# Called with the results returned from the Spider, after it has processed the response.
# Must return an iterable of Request, or item objects.
        for i in result:
            yield i
    def process_spider_exception(self, response, exception, spider):
# Called when a spider or process_spider_input() method (from other spider middleware) raises an exception.
# Should return either None or an iterable of Request or item objects.
        pass
    def process_start_requests(self, start_requests, spider):
# Called with the start requests of the spider, and works
# similarly to the process_spider_output() method, except
# that it doesn’t have a response associated.
# Must return only requests (not items).
        for r in start_requests:
            yield r
    def spider_opened(self, spider):
        spider.logger.info('Spider opened: %s' % spider.name)

class A2ADownloaderMiddleware:
# Not all methods need to be defined. If a method is not defined,
# scrapy acts as if the downloader middleware does not modify the
# passed objects.
    @classmethod
    def from_crawler(cls, crawler):
        # This method is used by Scrapy to create your spiders.
        s= cls()
        crawler.signals.connect(s.spider_opened, signal=signals.spider_opened)
        return s
    def process_request(self, request, spider):
       # Called for each request that goes through the downloader middleware.
# Must either:
# - return None: continue processing this request
# - or return a Response object
# - or return a Request object
# - or raise IgnoreRequest: process_exception() methods of installed downloader middleware will be called
        return None
    def process_response(self, request, response, spider):
# Called with the response returned from the downloader.
# Must either;
# - return a Response object
# - return a Request object
# - or raise IgnoreRequest
        return response
    def process_exception(self, request, exception, spider):
# Called when a download handler or a process_request()
# (from other downloader middleware) raises an exception.
# Must either:
# - return None: continue processing this exception
# - return a Response object: stops process_exception() chain
# - return a Request object: stops process_exception() chain
        pass
    def spider_opened(self, spider):
        spider.logger.info('Spider opened: %s' % spider.name)
```

Settings.py

```
# Scrapy settings for *NameOfTheProject* project
# Eliminate the hashtags to allow the settings
# For simplicity, this file contains only settings considered important or
# commonly used. You can find more settings consulting the documentation:
#
#     https://docs.scrapy.org/en/latest/topics/settings.html
#     https://docs.scrapy.org/en/latest/topics/downloader-middleware.html
#     https://docs.scrapy.org/en/latest/topics/spider-middleware.html

BOT_NAME = '*NameOfTheProject*'
SPIDER_MODULES = ['*NameOfTheProject*.spiders']
NEWSPIDER_MODULE = '*NameOfTheProject*.spiders'

# Crawl responsibly by identifying yourself (and your website) on the user-agent
#USER_AGENT = '*NameOfTheProject* (+http://www.yourdomain.com)'

# Obey robots.txt rules
#ROBOTSTXT_OBEY = False

# Configure maximum concurrent requests performed by Scrapy (default: 16)
#CONCURRENT_REQUESTS = 32

# Configure a delay for requests for the same website (default: 0)
# See https://docs.scrapy.org/en/latest/topics/settings.html#download-delay
# See also autothrottle settings and docs
#DOWNLOAD_DELAY = 2.1
# The download delay setting will honor only one of:
#CONCURRENT_REQUESTS_PER_DOMAIN = 16
#CONCURRENT_REQUESTS_PER_IP = 16

# Disable cookies (enabled by default)
#COOKIES_ENABLED = False

# Disable Telnet Console (enabled by default)
#TELNETCONSOLE_ENABLED = False

# Override the default request headers:
#DEFAULT_REQUEST_HEADERS = {
#   'Accept': #'text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8',
#   'Accept-Language': 'en',
#}

# Enable or disable spider middlewares
# See https://docs.scrapy.org/en/latest/topics/spider-middleware.html
#SPIDER_MIDDLEWARES = {
#    '*NameOfTheProject*.middlewares.A2ASpiderMiddleware': 543,
#}

# Enable or disable downloader middlewares
# See https://docs.scrapy.org/en/latest/topics/downloader-middleware.html
#DOWNLOADER_MIDDLEWARES = {
#    '*NameOfTheProject*.middlewares. *NameOfTheProject* #DownloaderMiddleware': 543,
#}

# Enable or disable extensions
# See https://docs.scrapy.org/en/latest/topics/extensions.html
#EXTENSIONS = {
#    'scrapy.extensions.telnet.TelnetConsole': None,
#}

# Configure item pipelines
# See https://docs.scrapy.org/en/latest/topics/item-pipeline.html
#ITEM_PIPELINES = {
#    '*NameOfTheProject*.pipelines. *NameOfTheProject* Pipeline': 300,
#}

# Enable and configure the AutoThrottle extension (disabled by default)
# See https://docs.scrapy.org/en/latest/topics/autothrottle.html
#AUTOTHROTTLE_ENABLED = True
# The initial download delay
#AUTOTHROTTLE_START_DELAY = 5
# The maximum download delay to be set in case of high latencies
#AUTOTHROTTLE_MAX_DELAY = 60
# The average number of requests Scrapy should be sending in parallel to
# each remote server
#AUTOTHROTTLE_TARGET_CONCURRENCY = 1.0
# Enable showing throttling stats for every response received:
#AUTOTHROTTLE_DEBUG = False

# Enable and configure HTTP caching (disabled by default)
# See https://docs.scrapy.org/en/latest/topics/downloader-middleware.html#httpcache-middleware-settings
#HTTPCACHE_ENABLED = True
#HTTPCACHE_EXPIRATION_SECS = 0
#HTTPCACHE_DIR = 'httpcache'
#HTTPCACHE_IGNORE_HTTP_CODES = []
#HTTPCACHE_STORAGE = 'scrapy.extensions.httpcache.FilesystemCacheStorage'
```

Spider.py code for each company:

A2A

```
import scrapy
from scrapy import Spider
from scrapy.http import Request
from scrapy.spiders import CrawlSpider, Rule
from scrapy.linkextractors import LinkExtractor

class A2aSpiderSpider(CrawlSpider):
    name = 'A2A_spider'
    allowed_domains = ['a2a.eu']
    start_urls = (
        'http://a2a.eu/',
    )
    rules = (Rule(LinkExtractor(allow=('sostenibilita')), callback='parse_text', follow = True),)

    def parse_text(self, response):
        text = response.xpath('//p').extract()
        yield{'text': text,
            }
        pass
```

ACEA

```
import scrapy
from scrapy import Spider
from scrapy.http import Request
from scrapy.spiders import CrawlSpider, Rule
from scrapy.linkextractors import LinkExtractor

class AceaSpiderSpider(CrawlSpider):
    name = 'ACEA_spider'
    allowed_domains = ['acea.it']
    start_urls = (
                'https://gruppo.acea.it/',
    )
    rules = (Rule(LinkExtractor(allow=('il-nostro-impegno')), callback='parse_text', follow = True),)

    def parse_text(self, response):
        text = response.xpath('//p').extract()
        yield{'text': text,
            }
        pass
```

ACSM-AGAM

```
import scrapy
from scrapy import Spider
from scrapy.http import Request
from scrapy.spiders import CrawlSpider, Rule
from scrapy.linkextractors import LinkExtractor

class AcsmSpiderSpider(scrapy.Spider):
    name = 'ACSM_spider'
    allowed_domains = ['acsm-agam.it']
    start_urls = (
            'http://www.acsm-agam.it/profilo',
    )
    def parse(self, response):
        text = response.xpath('//p').extract()
        yield{'text': text,
            }
        pass
```

Aeroporto di Bologna

```
import scrapy
from scrapy import Spider
from scrapy.http import Request
from scrapy.spiders import CrawlSpider, Rule
from scrapy.linkextractors import LinkExtractor



class AeroboloSpiderSpider(CrawlSpider):
    name = 'AeroBolo_spider'
    allowed_domains = ['bologna-airport.it']
    start_urls = ['https://www.bologna-airport.it/la-societa/ambiente-qualita-e-sicurezza/il-mondo-e-
una-porta-aperta/?idC=62593']
    rules = (Rule(LinkExtractor(allow=('ambiente-qualita-e-sicurezza')), callback='parse_text', follow =
True),)

    def parse_text(self, response):
        text = response.xpath('//p').extract()
        yield{'text': text,
            }
        pass
```

ALERION CLEAN ENERGY

```
import scrapy
from scrapy import Spider
from scrapy.http import Request

class AlerionSpiderSpider(scrapy.Spider):
    name = 'ALERION_spider'
    allowed_domains = ['alerion.it']
    def start_requests(self):
        urls = [
            'http://alerion.it/wind/#13/',
        ]
        for url in urls:
            yield scrapy.Request(url=url, callback = self.parse)

    def parse(self, response):
        text = response.xpath('//p').extract()
        yield{'text': text,
            }
        pass
```

AMPLIFON

```
import scrapy
from scrapy import Spider
from scrapy.http import Request
from scrapy.spiders import CrawlSpider, Rule
from scrapy.linkextractors import LinkExtractor

class AmpliSpiderSpider(CrawlSpider):
    name = 'AMPLI_spider'
    allowed_domains = ['corporate.amplifon.com']
    start_urls = ['http://corporate.amplifon.com/it/sostenibilita']

    rules = (Rule(LinkExtractor(allow=('sostenibilita')), callback='parse_text', follow = True),)

    def parse_text(self, response):
        text = response.xpath('//p').extract()
        yield{'text': text,
            }
        pass
```

ATLANTIA

```
import scrapy
from scrapy import Spider
from scrapy.http import Request
from scrapy.spiders import CrawlSpider, Rule
from scrapy.linkextractors import LinkExtractor

class AtlantiaSpiderSpider(CrawlSpider):
    name = 'ATLANTIA_spider'
    allowed_domains = ['atlantia.it']
    start_urls = ['http://www.atlantia.it/it/sostenibilita/']

    rules = (Rule(LinkExtractor(allow=('sostenibilita')), callback='parse_text', follow = True),)

    def parse_text(self, response):
        text = response.xpath('//p').extract()
        yield{'text': text,
            }
        pass
```

CAMPARI

```
import scrapy
from scrapy import Spider
from scrapy.http import Request
from scrapy.spiders import CrawlSpider, Rule
from scrapy.linkextractors import LinkExtractor

class CampariSpiderSpider(CrawlSpider):
   name = 'CAMPARI_spider'
   allowed_domains = ['camparigroup.com']
   start_urls = ['http://camparigroup.com/it/page/sostenibilita/']

   rules = (Rule(LinkExtractor(allow=('sostenibilita')), callback='parse_text', follow = True),)

   def parse_text(self, response):
      text = response.xpath('//p').extract()
      yield{'text': text,
         }
      pass
```

CNH

```
import scrapy
from scrapy import Spider
from scrapy.http import Request
from scrapy.spiders import CrawlSpider, Rule
from scrapy.linkextractors import LinkExtractor


class CnhSpiderSpider(CrawlSpider):
   name = 'CNH_spider'
   allowed_domains = ['cnhindustrial.com']
   start_urls = (
            'https://www.cnhindustrial.com/it-
it/sustainability/our_approach_to_sustainability/Pages/default.aspx',
            )
   rules = (Rule(LinkExtractor(allow=('sustainability')), callback='parse_text', follow = True),)

   def parse_text(self, response):
      text = response.xpath('//p').extract()
      yield{'text': text,
         }
      pass
```

DIASORIN

```
import scrapy
from scrapy import Spider
from scrapy.http import Request
from scrapy.spiders import CrawlSpider, Rule
from scrapy.linkextractors import LinkExtractor

class DiasorinSpiderSpider(CrawlSpider):
   name = 'DIASORIN_spider'
   allowed_domains = ['diasoringroup.com']
   start_urls = ['http://diasoringroup.com/it/sostenibilita/il-nostro-impegno']

   rules = (Rule(LinkExtractor(allow=('sostenibilita')), callback='parse_text', follow = True),)
```

```
    def parse_text(self, response):
        text = response.xpath('//p/text()').extract()
        more_text = response.xpath('//div/*[@class="field-item even"]/text()').extract()
        yield{
        'text': text,
        'more_text': more_text
            }
        pass
```

EDISON

```
import scrapy

class EdisonSpiderSpider(scrapy.Spider):
    name = 'EDISON_spider'
    allowed_domains = ['edison.it/it']
    start_urls = ('http://https://www.edison.it/it/sostenibilita/',
                'https://www.edison.it/it/impegno-edison-sviluppo-sostenibile',
                'https://www.edison.it/it/governance-sostenibilita',
                'https://www.edison.it/it/sostenibilita-processi-aziendali',
                'https://www.edison.it/it/lotta-al-cambiamento-climatico',
                'https://www.edison.it/it/risorse-umane',
                'https://www.edison.it/it/stakeholder-consumatori',
                'https://www.edison.it/it/dialogo-con-i-consumatori',
                'https://www.edison.it/it/sostenibilita-biodiversita',
                'https://www.edison.it/it/percorso-turistico-naturalistico-tracciolino',
                'https://www.edison.it/it/valore-sociale-stakeholder',
                'https://www.edison.it/it/salone-csr-innovazione-sociale',
                'https://www.edison.it/it/valorizziamo-il-talento-sul-territorio',
    )
    def parse(self, response):
        text = response.xpath('//p').extract()
        yield{'text': text,
            }
        pass
```

ELETTRA ENERGIA

```
import scrapy

class ElettraSpiderSpider(scrapy.Spider):
    name = 'ELETTRA_spider'
    allowed_domains = ['elettrainvestimenti.it']

    def start_requests(self):
        urls = [
            'https://www.elettrainvestimenti.it/sostenibilita-2/',
        ]
        for url in urls:
            yield scrapy.Request(url=url, callback = self.parse)

    def parse(self, response):
        text = response.xpath('//div/*[@class="desc"]/text()').extract()
        yield{'text': text,
            }
        pass
```

ENEL

```
import scrapy
from scrapy import Spider
from scrapy.http import Request
from scrapy.spiders import CrawlSpider, Rule
from scrapy.linkextractors import LinkExtractor

class EnelSpiderSpider(CrawlSpider):
    name = 'ENEL_spider'
    allowed_domains = ['enel.com']
    start_urls = (
            'https://www.enel.com/it/investitori/sostenibilita',
    )
    rules = (Rule(LinkExtractor(allow=('sostenibilita')), callback='parse_text', follow = True),)

    def parse_text(self, response):
        text = response.xpath('//p').extract()
        yield{'text': text,
            }
        pass
```

ERG

```
import scrapy
from scrapy import Spider
from scrapy.http import Request
from scrapy.spiders import CrawlSpider, Rule
from scrapy.linkextractors import LinkExtractor

class ErgSpiderSpider(CrawlSpider):
    name = 'ERG_spider'
    allowed_domains = ['erg.eu']
    start_urls = (
        'https://erg.eu/',
    )
    rules = (Rule(LinkExtractor(allow=('sostenibilita')), callback='parse_text', follow = True),)

    def parse_text(self, response):
        text = response.xpath('//*[@class="text-content"]').extract_first()
        text_expanded = response.xpath('//*[@class="accordion-item-text"]').extract()
        yield{'text': text,
            'text_expanded': text_expanded
            }
        pass
```

FALCK RENEWABLES

```

import scrapy



class FalckSpiderSpider(scrapy.Spider):
    name = 'FALCK_spider'
    allowed_domains = ['falckrenewables.com']
    start_urls = (
            'https://www.falckrenewables.com/communities',
    )

    def parse(self, response):
        text = response.xpath('//p').extract()
        yield{'text': text}
        pass
```

FERRARI

```
import scrapy
from scrapy import Spider
from scrapy.http import Request
from scrapy.spiders import CrawlSpider, Rule
from scrapy.linkextractors import LinkExtractor

class FerrariSpiderSpider(CrawlSpider):
    name = 'FERRARI_spider'
    allowed_domains = ['corporate.ferrari.com']
    start_urls = (
                'http://corporate.ferrari.com/it/chi-siamo/sostenibilita/',
                )
    rules = (Rule(LinkExtractor(allow=('sostenibilita')), callback='parse_text', follow = True),)

    def parse_text(self, response):
        text = response.xpath('//p').extract()
        yield{'text': text,
            }
        pass
```

FRENDY ENERGY

```
import scrapy
from scrapy import Spider
from scrapy.http import Request

class FrendySpiderSpider(scrapy.Spider):
    name = 'FRENDY_spider'
    allowed_domains = ['frendyenergy.edison.it']
    start_urls = ['https://frendyenergy.edison.it/la-societa/']

    def parse(self, response):
                text = response.xpath('//p').extract()
                yield{"text": text}
    pass
```

GENERALI

```
import scrapy
from scrapy import Spider
from scrapy.http import Request
from scrapy.spiders import CrawlSpider, Rule
from scrapy.linkextractors import LinkExtractor


class GeneraliSpiderSpider(CrawlSpider):
    name = 'GENERALI_spider'
    allowed_domains = ['generali.it']
    start_urls = (
                'https://www.generali.it/chi-siamo/sostenibilita/',
                )
    rules = (Rule(LinkExtractor(allow=('sostenibilita')), callback='parse_text', follow = True),)

    def parse_text(self, response):
        text = response.xpath('//p').extract()
        more_text = response.xpath('//li').extract()
        yield{
        'text': text,
        'more_text' : more_text,
            }
        pass
```

HERA

```
import scrapy
from scrapy import Spider
from scrapy.http import Request
from scrapy.spiders import CrawlSpider, Rule
from scrapy.linkextractors import LinkExtractor

class HeraSpiderSpider(CrawlSpider):
    name = 'HERA_spider'
    allowed_domains = ['gruppohera.it']
    start_urls = (
            'https://www.gruppohera.it/',
    )
    rules = (Rule(LinkExtractor(allow=('gruppo/responsabilita_sociale'),
deny=('responsabilita_sociale/bs/', 'vedo_hera/','news/')), callback='parse_text', follow = True),)

    def parse_text(self, response):
        text = response.xpath('//p').extract()
        yield{'text': text,
            }
        pass
```

INIZIATIVE BRESCIANE

```
import scrapy

class IniziativeSpiderSpider(scrapy.Spider):
    name = 'INIZIATIVE_spider'
    allowed_domains = ['iniziativebrescianespa.it']
    start_urls = ['http://www.iniziativebrescianespa.it/qualita_e_ambiente']

    def parse(self, response):
            text = response.xpath('//*[@class="MsoNormal"]/text()').extract()
            yield{"text": text}
    pass
```

INTESA SANPAOLO

```
import scrapy
from scrapy import Spider
from scrapy.http import Request
from scrapy.spiders import CrawlSpider, Rule
from scrapy.linkextractors import LinkExtractor

class IntesaSpiderSpider(CrawlSpider):
    name = 'INTESA_spider'
    allowed_domains = ['group.intesasanpaolo.com']
    start_urls = (
            'http://group.intesasanpaolo.com/it/sostenibilita',
            )
    rules = (Rule(LinkExtractor(allow=('sostenibilita')), callback='parse_text', follow = True),)

    def parse_text(self, response):
        text = response.xpath('//p').extract()
        yield{'text': text,
            }
        pass
```

IREN

```
import scrapy
from scrapy import Spider
from scrapy.http import Request
from scrapy.spiders import CrawlSpider, Rule
from scrapy.linkextractors import LinkExtractor


class IrenSpiderSpider(CrawlSpider):
    name = 'IREN_spider'
    allowed_domains = ['gruppoiren.it']
    start_urls = (
            'https://gruppoiren.it/',
    )
    rules = (Rule(LinkExtractor(allow=('sostenibilita', 'governance-della-sostenibilita', 'strumenti-di-csr',
            'matrice-di-materialita', 'contributo-agli-sdgs', 'servizi-sostenibili')), callback='parse_text',
follow = True),)

    def parse_text(self, response):
        text = response.xpath('//p').extract()
        yield{'text': text,
            }
        pass
```

ITALGAS

```
import scrapy


from scrapy import Spider
from scrapy.http import Request
from scrapy.spiders import CrawlSpider, Rule
from scrapy.linkextractors import LinkExtractor

class ItalgasSpiderSpider(CrawlSpider):
    name = 'ITALGAS_spider'
    allowed_domains = ['italgas.it']
    start_urls = ['https://www.italgas.it/it/il-nostro-impegno/']

    rules = (Rule(LinkExtractor(allow=('il-nostro-impegno')), callback='parse_text', follow = True),)

    def parse_text(self, response):
        text = response.xpath('//p').extract()
        yield{'text': text,
            }
        pass
```

MONCLER

```
import scrapy
from scrapy import Spider
from scrapy.http import Request
from scrapy.spiders import CrawlSpider, Rule
from scrapy.linkextractors import LinkExtractor

class MonclerSpiderSpider(CrawlSpider):
    name = 'MONCLER_spider'
    allowed_domains = ['monclergroup.com']
    start_urls = ['https://www.monclergroup.com/it/sostenibilita/']

    rules = (Rule(LinkExtractor(allow=('sostenibilita')), callback='parse_text', follow = True),)

    def parse_text(self, response):
        text = response.xpath('//p').extract()
        yield{'text': text,
            }
        pass
```

NEXI

```
import scrapy
from scrapy import Spider
from scrapy.http import Request
from scrapy.spiders import CrawlSpider, Rule
from scrapy.linkextractors import LinkExtractor



class NexiSpiderSpider(CrawlSpider):
    name = 'NEXI_spider'
    allowed_domains = ['nexi.it']
    start_urls = ['http://nexi.it/chi-siamo/brand/sostenibilita.html']

    rules = (Rule(LinkExtractor(allow=('sostenibilita')), callback='parse_text', follow = True),)

    def parse_text(self, response):
        text = response.xpath('//p').extract()
        yield{'text': text,
            }
        pass
```

POSTE ITALIANE

```
import scrapy

class PosteSpiderSpider(scrapy.Spider):
    name = 'POSTE_spider'
    allowed_domains = ['posteitaliane.it']
    start_urls = (
            'https://www.posteitaliane.it/it/il-nostro-approccio.html',
            'https://www.posteitaliane.it/it/i-nostri-stakeholder.html',
            'https://www.posteitaliane.it/it/matrice-di-materialita.html',
            'https://www.posteitaliane.it/it/matrice-impatto-poste-italiane.html',
            'https://www.posteitaliane.it/it/interconnessione-temi-materiali-metriche.html',
            'https://www.posteitaliane.it/it/impatti-generati-poste-italiane.html',
            'https://www.posteitaliane.it/it/valore-economico-generato-distribuito.html',
            'https://www.posteitaliane.it/it/integrita-trasparenza.html',
            'https://www.posteitaliane.it/it/percorso-sostenibilita-poste-italiane.html',
        'https://www.posteitaliane.it/it/deliver-2022.html',
        'https://www.posteitaliane.it/it/strategia-modello-business.html',
        'https://www.posteitaliane.it/it/politiche-di-sostenibilita.html',
        'https://www.posteitaliane.it/it/piano-strategico-esg-gruppo.html',
        'https://www.posteitaliane.it/it/presentazione-bilancio.html#/video',
        'https://www.posteitaliane.it/it/politica-integrata-del-gruppo-poste-italiane.html',
        'https://www.posteitaliane.it/it/tappe-percorso-integrita.html',
        'https://www.posteitaliane.it/it/sostenibilita-fornitori.html',
        'https://www.posteitaliane.it/it/compliance-per-la-tutela-della-concorrenza-e-del-
consumatore.html',
        'https://www.posteitaliane.it/it/valorizzazione-delle-persone.html',
        'https://www.posteitaliane.it/it/formazione-sviluppo.html',
        'https://www.posteitaliane.it/it/welfare-benessere-personale.html',
        'https://www.posteitaliane.it/it/volontariato-di-impresa.html',
        'https://www.posteitaliane.it/it/salute-sicurezza.html',
        'https://www.posteitaliane.it/it/valore-poste-ascolta-il-personale.html',
        'https://www.posteitaliane.it/it/relazioni-parti-sociali.html',
        'https://www.posteitaliane.it/it/diversita-inclusione.html',
        'https://www.posteitaliane.it/it/politica-in-materia-di-diritti-umani.html',
        'https://www.posteitaliane.it/it/politica-in-materia-di-diversit-e-inclusione.html',
        'https://www.posteitaliane.it/it/pari-opportunita.html',
        'https://www.posteitaliane.it/it/sostegno-al-territorio.html',
        'https://www.posteitaliane.it/it/dialogo-trasparenza-istituzioni.html',
        'https://www.posteitaliane.it/it/politica-iniziative-comunita.html',
        'https://www.posteitaliane.it/it/inclusione-finanziaria.html',
        'https://www.posteitaliane.it/it/customer-experience.html',
        'https://www.posteitaliane.it/it/consumatori.html',
        'https://www.posteitaliane.it/it/dialogando-consumatori.html',
        'https://www.posteitaliane.it/it/educazione-finanziaria.html',
        'https://www.posteitaliane.it/it/sicurezza-informatica.html',
        'https://www.posteitaliane.it/it/sostenibilita-innovazione.html',
        'https://www.posteitaliane.it/it/innovazione-digitalizzazione-processi.html',
        'https://www.posteitaliane.it/it/innovazione-digitalizzazione-prodotti-servizi.html',
        'https://www.posteitaliane.it/it/decarbonizzazione-immobili-logistica.html',
        'https://www.posteitaliane.it/it/politica-sostenibilita-ambientale.html',
        'https://www.posteitaliane.it/it/impatti-ambientali-immobili.html',
        'https://www.posteitaliane.it/it/impatti-ambientali-logistica.html',
        'https://www.posteitaliane.it/it/gestione-rendicontazione-rischi-cambiamento-
climatico.html',
        'https://www.posteitaliane.it/it/progetto-platoon.html',
        'https://www.posteitaliane.it/it/integrazione-esg-investimento.html',
        'https://www.posteitaliane.it/it/integrazione-esg-politiche-di-assicurazione.html',
        'https://www.posteitaliane.it/it/esg-informazioni.html',
  )
  def parse(self, response):
    text = response.xpath('//div/*[@class="box-editable-area box-editable-
spacing"]/text()').extract()
    text_plus = response.xpath('//div/*[@class="panel-body"]/text()').extract()
    more_text = response.xpath('//td').extract()
    first_text = response.xpath('//small/*[@class= "block"]/text()').extract()
    yield{
      'text': text,
      'text_plus': text_plus,
      'more_text': more_text,
      'fist_text': first_text,
      }
    pass
```

RENERGETICA

```
import scrapy


class RenergeticaSpiderSpider(scrapy.Spider):
  name = 'RENERGETICA_spider'
  allowed_domains = ['renergetica.com']
  start_urls = ['https://www.renergetica.com/reti-ibride/']
  def parse(self, response):
    text = response.xpath('//p').extract()
    yield{'text': text,
      }
    pass
```

SNAM

```
import scrapy


class SnamSpiderSpider(scrapy.Spider):
  name = 'SNAM_spider'
  allowed_domains = ['snam.it']
  start_urls = ('https://www.snam.it/it/sostenibilita/',
'https://www.snam.it/it/sostenibilita/strategia_per_futuro/',
'https://www.snam.it/it/sostenibilita/strategia_per_futuro/snam_net_zero_carbon.html',
        'https://www.snam.it/it/sostenibilita/strategia_per_futuro/esg_scorecard.html',
        'https://www.snam.it/it/sostenibilita/impegni_snam/',
        'https://www.snam.it/it/sostenibilita/impegni_snam/analisi_di_materialita.html',
        'https://www.snam.it/it/sostenibilita/impegni_snam/il_modello_snam.html',
        'https://www.snam.it/it/sostenibilita/impegni_snam/valore_condiviso.html',
        'https://www.snam.it/it/sostenibilita/impegni_snam/dialogo_con_gli_stakeholder.html',
        'https://www.snam.it/it/sostenibilita/impegni_snam/task_force_CFD.html',
        'https://www.snam.it/it/sostenibilita/impegni_snam/snam_nel_global_compact.html',
        'https://www.snam.it/it/sostenibilita/impegni_snam/etica_d_impresa_e_governance.html',
        'https://www.snam.it/it/sostenibilita/responsabilita_verso_tutti/',

        'https://www.snam.it/it/sostenibilita/responsabilita_verso_tutti/utilizzo_del_gas_naturale.h
tml',

        'https://www.snam.it/it/sostenibilita/responsabilita_verso_tutti/sostenibilita_delle_infrastr
utture.html',
        'https://www.snam.it/it/sostenibilita/responsabilita_verso_tutti/generare_valore.html',

        'https://www.snam.it/it/sostenibilita/responsabilita_verso_tutti/continuita_del_servizio.ht
ml',

        'https://www.snam.it/it/sostenibilita/responsabilita_verso_tutti/uso_sicuro_degli_impianti_
a_gas.html',
        'https://www.snam.it/it/sostenibilita/responsabilita_verso_tutti/Impegno_sociale.html',

        'https://www.snam.it/it/sostenibilita/responsabilita_verso_tutti/ritrovamenti_archeologici.h
tml',
        'https://www.snam.it/it/sostenibilita/responsabilita_verso_tutti/fondazione.html',
        'https://www.snam.it/it/sostenibilita/agire_per_ambiente/',
        'https://www.snam.it/it/sostenibilita/agire_per_ambiente/sistemi_di_gestione_ambientali.h
tml',
        'https://www.snam.it/it/sostenibilita/agire_per_ambiente/climate_change.html',
        'https://www.snam.it/it/sostenibilita/agire_per_ambiente/consumi_energetici.html',
        'https://www.snam.it/it/sostenibilita/agire_per_ambiente/aria.html',
        'https://www.snam.it/it/sostenibilita/agire_per_ambiente/fonti_rinnovabili.html',
        'https://www.snam.it/it/sostenibilita/agire_per_ambiente/biodiversita.html',
        'https://www.snam.it/it/sostenibilita/agire_per_ambiente/acqua.html',
        'https://www.snam.it/it/sostenibilita/agire_per_ambiente/rifiuti.html',
        'https://www.snam.it/it/sostenibilita/lavorare_in_sicurezza/',
        'https://www.snam.it/it/sostenibilita/lavorare_in_sicurezza/una_cultura_di_sicurezza.html',

        'https://www.snam.it/it/sostenibilita/lavorare_in_sicurezza/il_ruolo_dei_fornitori_e_appalt
atori.html',
        'https://www.snam.it/it/sostenibilita/crescere_con_fornitori/',

        'https://www.snam.it/it/sostenibilita/crescere_con_fornitori/la_sostenibilita_dei_fornitori.h
tml',

        'https://www.snam.it/it/sostenibilita/crescere_con_fornitori/il_valore_della_partnership.ht
ml',
        'https://www.snam.it/it/sostenibilita/valorizzare_le_persone/',

        'https://www.snam.it/it/sostenibilita/valorizzare_le_persone/l_importanza_della_diversita.
html',

        'https://www.snam.it/it/sostenibilita/valorizzare_le_persone/la_valorizzazione_dei_talenti.h
tml',
        'https://www.snam.it/it/sostenibilita/valorizzare_le_persone/lavoratori_al_centro.html',

        'https://www.snam.it/it/sostenibilita/valorizzare_le_persone/iniziative_per_i_dipendenti.ht
ml',
        'https://www.snam.it/it/sostenibilita/reporting_e_performance/',

        'https://www.snam.it/it/sostenibilita/reporting_e_performance/il_report_sulla_responsabili
ta_sociale.html',
        'https://www.snam.it/it/sostenibilita/reporting_e_performance/bilancio_integrato.html',
        'https://www.snam.it/it/sostenibilita/reporting_e_performance/linee_guida_e_gri.html',
  )

  def parse(self, response):
    text = response.xpath('//p').extract()

    yield{'text': text,
      }
    Pass
```

TIM

```
import scrapy
from scrapy import Spider
from scrapy.http import Request
from scrapy.spiders import CrawlSpider, Rule
from scrapy.linkextractors import LinkExtractor

class TimSpiderSpider(CrawlSpider):
    name = 'TIM_spider'
    allowed_domains = ['gruppotim.it']
    start_urls = ['http://www.gruppotim.it/it/sostenibilita/strategia/futuro-sostenibile/modello.html/']

    rules = (Rule(LinkExtractor(allow=('sostenibilita')), callback='parse_text', follow = True),)

    def parse_text(self, response):
        text = response.xpath('//p').extract()
        yield{'text': text,
            }
        pass
```

UNICREDIT

```
import scrapy
from scrapy import Spider
from scrapy.http import Request
from scrapy.spiders import CrawlSpider, Rule
from scrapy.linkextractors import LinkExtractor

class UnicreditSpiderSpider(CrawlSpider):
    name = 'UNICREDIT_spider'
    allowed_domains = ['unicreditgroup.eu']
    start_urls = (
            'http://www.unicreditgroup.eu/it/a-sustainable-bank/',
            'https://www.unicreditgroup.eu/it/a-sustainable-bank.html?topmenu=INT-
TM_SUS00_it099',
            )

    rules = (Rule(LinkExtractor(allow=('a-sustainable-bank')), callback='parse_text', follow = True),)

    def parse_text(self, response):
        text = response.xpath('//p').extract()
        yield{'text': text,
            }
        pass
```

Appendix B

Text manipulation & word counting

It is presented the code to count the csv files words and to calculate the words matching with the created dictionary (the variable called “selected_words”).

```
import re
import operator
import tkinter as tk
from tkinter import filedialog
from tkinter.filedialog import askdirectory
from os import listdir
from os.path import isfile, join

selected_words = '''biodiversi, biodegrad, carboni, clim, monossid, deforest, desertif, siccit, terremot,
energi, alluvion, mar, fium, risors, natura, riscaldament, serra, rinnov, ozono, inquin, vulcan,
spazzatur, ambient, atmosfer, eco, scart, montagn, vent, sol, acqu, suolo, verd, minor, gener,
territori, cambiament, monous, salvaguard, estin, sprec, ricil, atten, scrat, combatt, conserv, prote,
evol, permess, preven, cambi, riform, sicur, ripristin,  benefic, salut, avanza, verific, soste, proced,
iniziativ, procedur, ottimiz, rinnov, sicur, standard, traspare, collabor, condivi, impegn, positiv,
innova, qualit, inclus, accord,  etic, diritt, dover, socioeconomic, civil, social, comunit, unit, riform,
istituzion, organizzazion, leader, soci, norma, equi, comuni, accord, contratt, comitat, responsabil,
maternit, paternit, management, consigli, amministr, iniziativ, cultur, person'''

def get_csv_files():
    #Open tab to select the folder with the .csv files
        root = tk.Tk()
        root.withdraw()
        file_path = askdirectory()

    #For each file in the folder if it has the last 4 letters
    #equal to .csv then select its path and the name
        onlyfiles = [f for f in listdir(file_path) if isfile(join(file_path, f)) and f[-4:] == '.csv']
        return file_path, onlyfiles


def get_all_words(filepath):
        print('Reading File %s' % filepath)
        # Open the file in read mode with latin1 since the text is in
    # italian
        text = open(filepath, "r", encoding='latin1')

        # Create an empty dictionary
        all_words = dict()

        # Loop through each line of the file
        for line in text:
      # re.sub used to substitute. re.sub(pattern,repl,string)
                line = re.sub('<.*?>', '', line)
                line = re.sub('"', '', line)
                line = re.sub('\.', '', line)

                # Remove the leading & trailing white spaces
                line = line.strip()

                # Convert the characters in line to
                # lowercase to avoid mismatch with my Dictionary
                line = line.lower()

                # Split the line into words
                words = line.split(" ")

                # Iterate over each word in line
                for word in words:
                    if word == "" or word == '-':
                        continue
     #if the word is not a number
                    if not word.isalnum():
                        continue

                    # Check if the word is already in dictionary
                    if word in all_words:
                        all_words[word] = all_words[word] + 1
                    else:
                        # Add the word to dictionary with count 1
                        all_words[word] = 1
    return all_words

def write_to_file(filename, words):
 #open file to write on it
    with open(filename, 'w') as f:
        for key in list(words.keys()):
            f.write(f"{key},{words[key]}\n")
def get_selected_words_dict(all_words, words_list):
    d = {}
    for word in words_list.split(', '):
        if word not in d:
            d[word] = 0


        for all_word in all_words:
            if word in all_word:
                d[word] += all_words.get(all_word, 0)
    d = dict(sorted(d.items(), key=operator.itemgetter(1), reverse=True)[:50000])  return d

filepath, files = get_csv_files()

if len(files) == 0:
    print('No csv files found')
    exit()

for f in files:
    fp = join(filepath, f)
    all_words = get_all_words(fp)

    top50 = dict(sorted(all_words.items(), key=operator.itemgetter(1), reverse=True)[:50000])
```

```
# Print the contents of dictionary
for key in list(top50.keys()):
    print(key, ":", top50[key])
print()

fn = f.replace('.csv', '_top50.csv')
write_to_file(fn, top50)
selected_words_dict = get_selected_words_dict(all_words, selected_words)
print(selected_words_dict)
fn2 = f.replace('.csv', '_selected.csv')
write_to_file(fn2, selected_words_dict)
```

Appendix C

Bar plotting of CSV files

The following code is used to plot the data of the CSV files containing the counted words.

To run the code, it is necessary to change the highlighted lines and insert in the first argument the directory path of the CSV file and in the second the name of the file to plot.

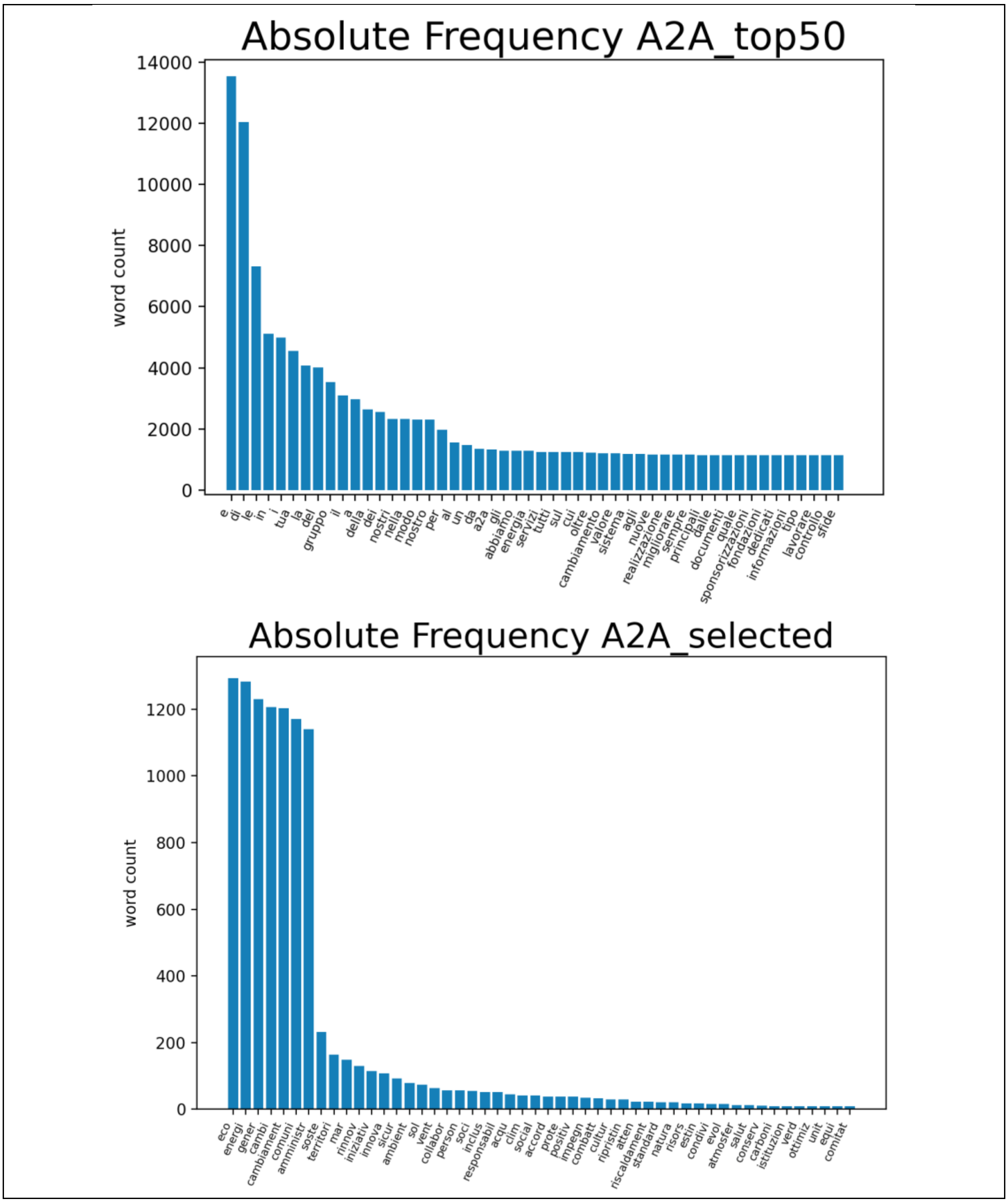
```
import pandas as pd
import matplotlib as mlp
import matplotlib.pyplot as plt
import os
os.chdir("insert_here_the_directory_of_the_csv_files")
name= ‘File_name.csv‘
dataframe = pd.read_csv(name, names=['word', 'absolute frequency’],header=None, index_col=0)

#creating a figure and the axes
fig, ax = plt.subplots()

dataframe = dataframe.sort_values('absolute frequency', ascending=False)
ax.bar(dataframe.index, dataframe['absolute frequency'])

#to make the x line (containing the words) better readable
ax.set_xticklabels(dataframe.index, rotation = 65, horizontalalignment = 'right', fontsize = '10')
ax.set_title('Absolute Frequency '+name[:-4], fontsize=22)
ax.set_ylabel('word count')
plt.show()
```

Appendix D

Bar charts of frequencies

In this section are presented the bar charts of the frequencies of the companies' sustainability words used on their websites and the frequency distribution of my dictionary's words for each company.

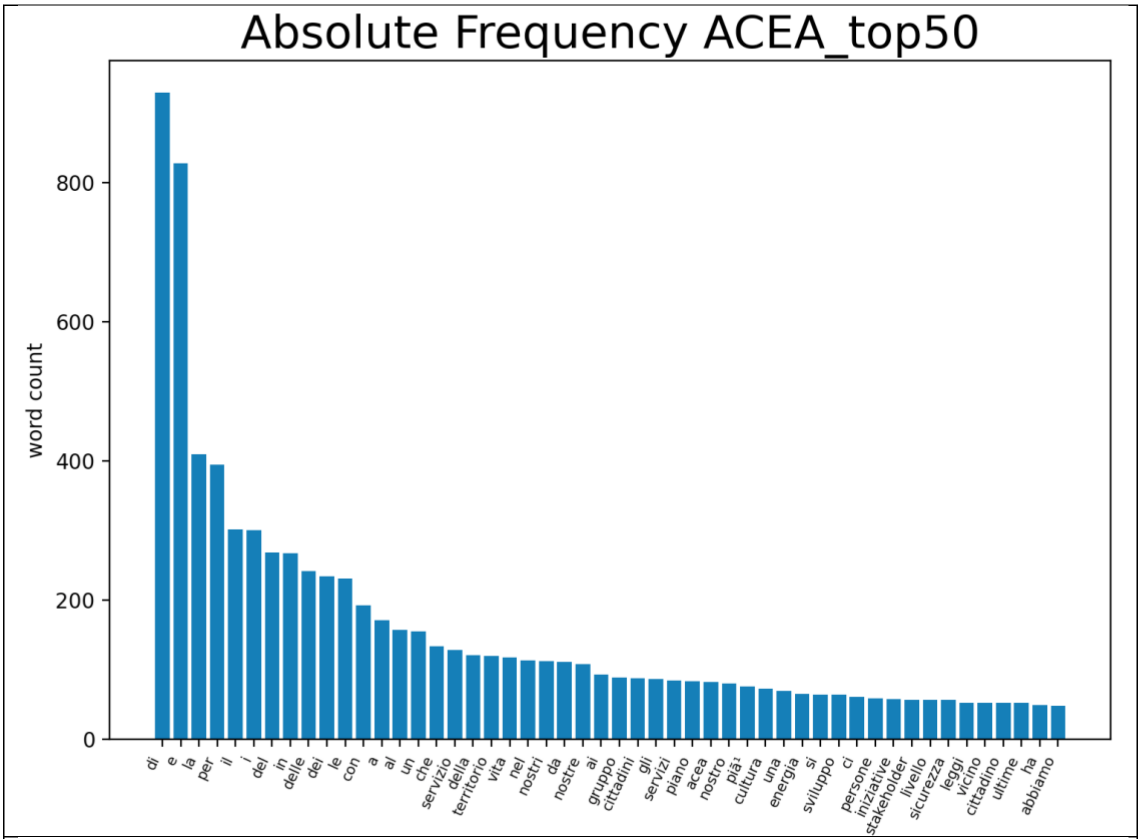
A2A

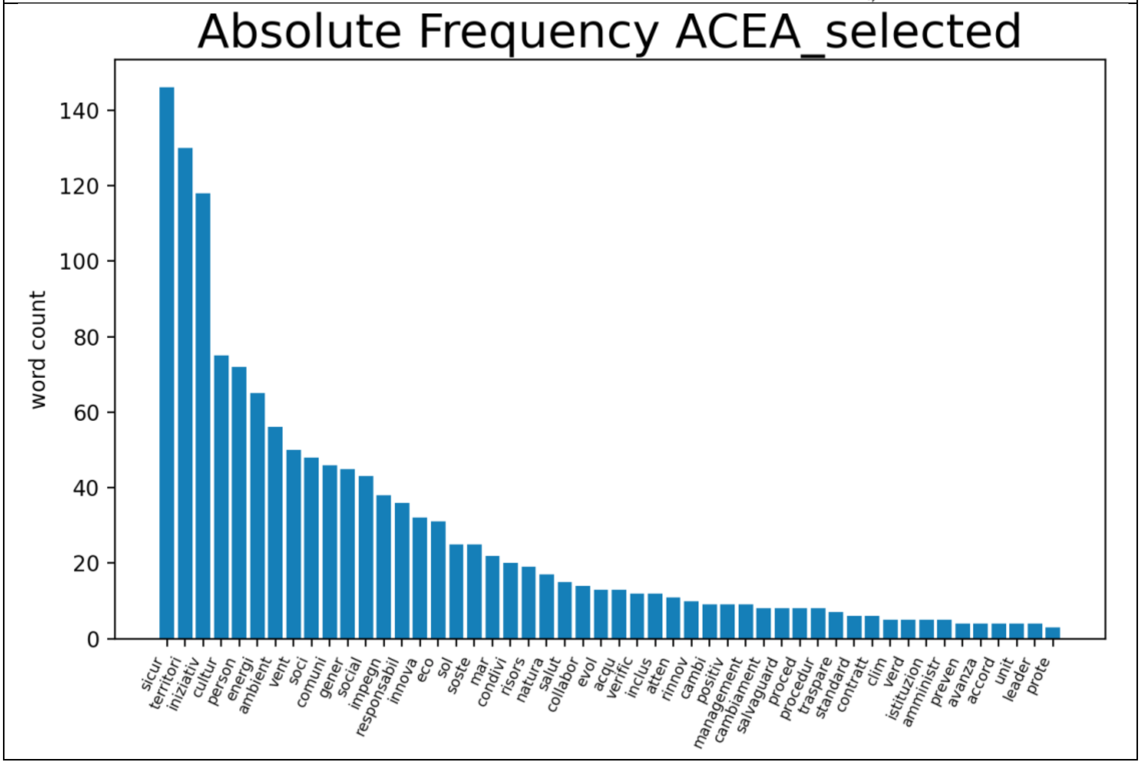


Absolute Frequency A2A_top50

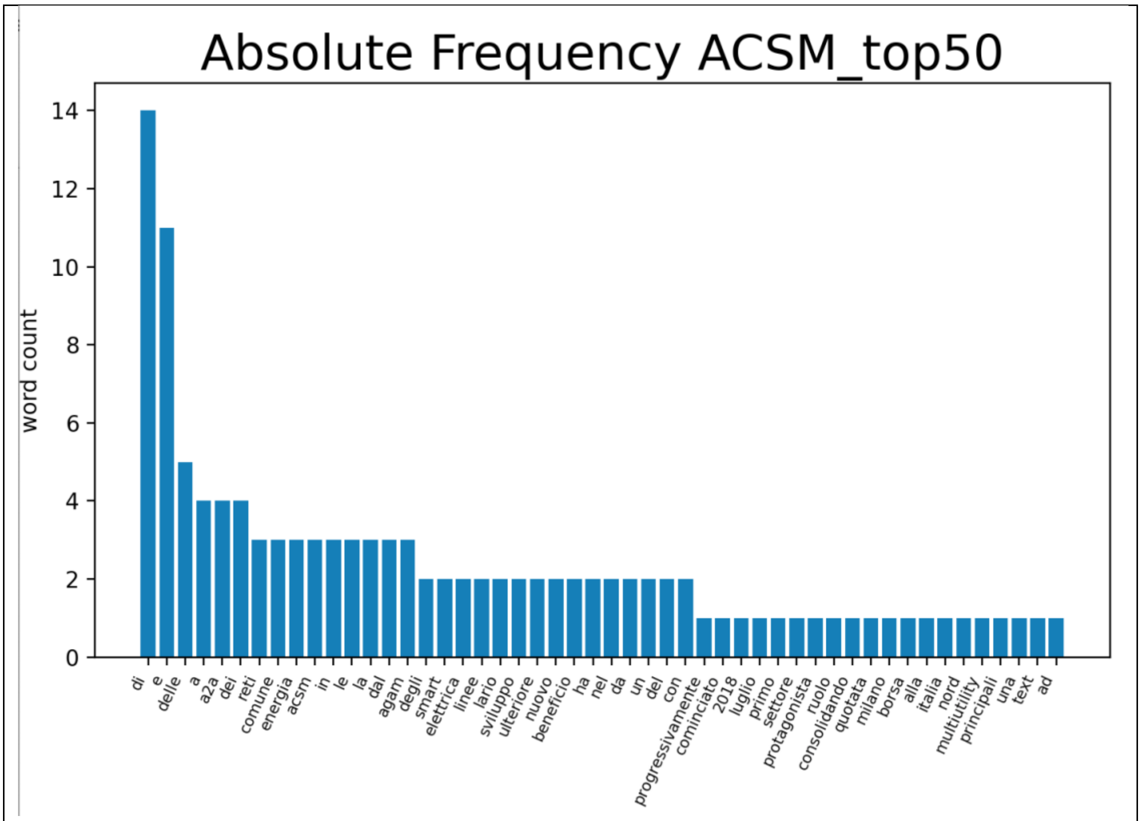
Absolute Frequency A2A_selected

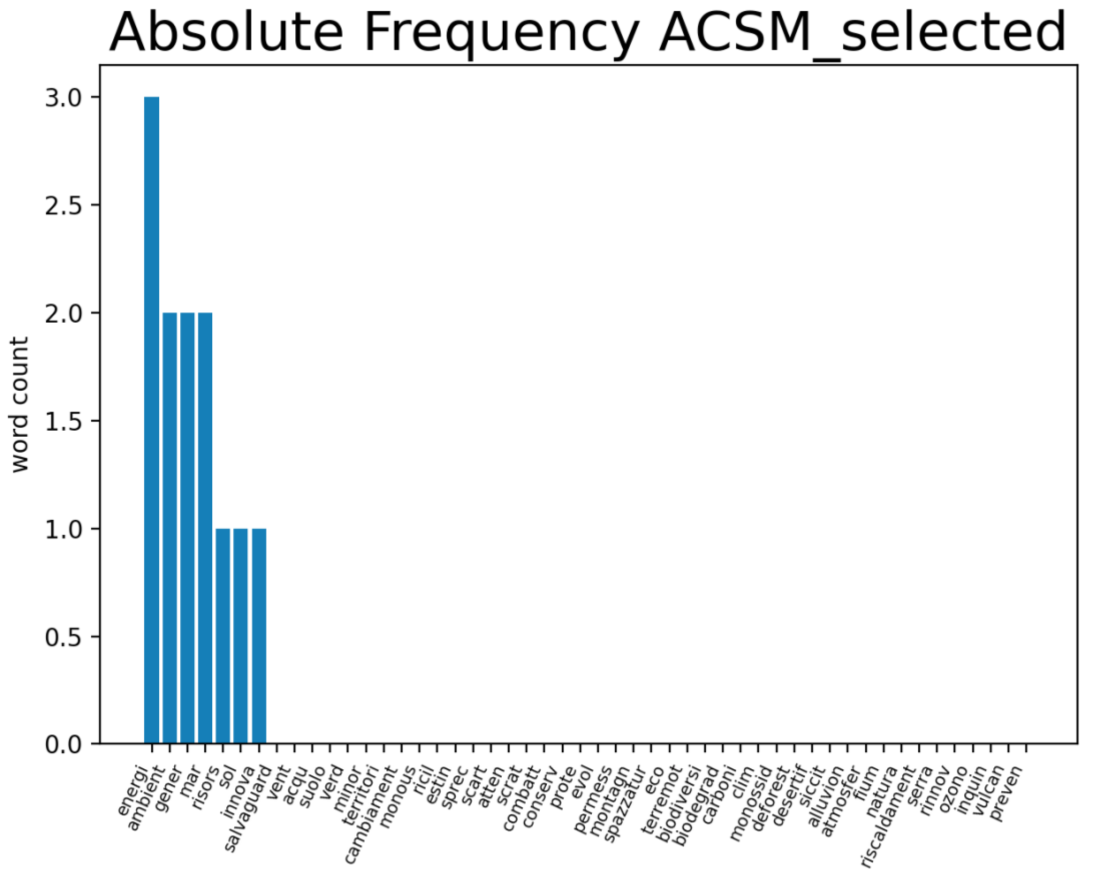
ACEA


Absolute Frequency ACEA_top50
word count
0
200
400
600
800
di
e
la
per
il
i
del
in
delle
dei
le
con
a
al
un
che
servizio
della
territorio
vita
nel
nostri
da
nostre
ai
gruppo
cittadini
gli
servizi
piano
acea
nostro
piã¹
cultura
una
energia
si
sviluppo
ci
persone
iniziative
stakeholder
livello
sicurezza
leggi
vicino
cittadino
ultime
ha
abbiamo

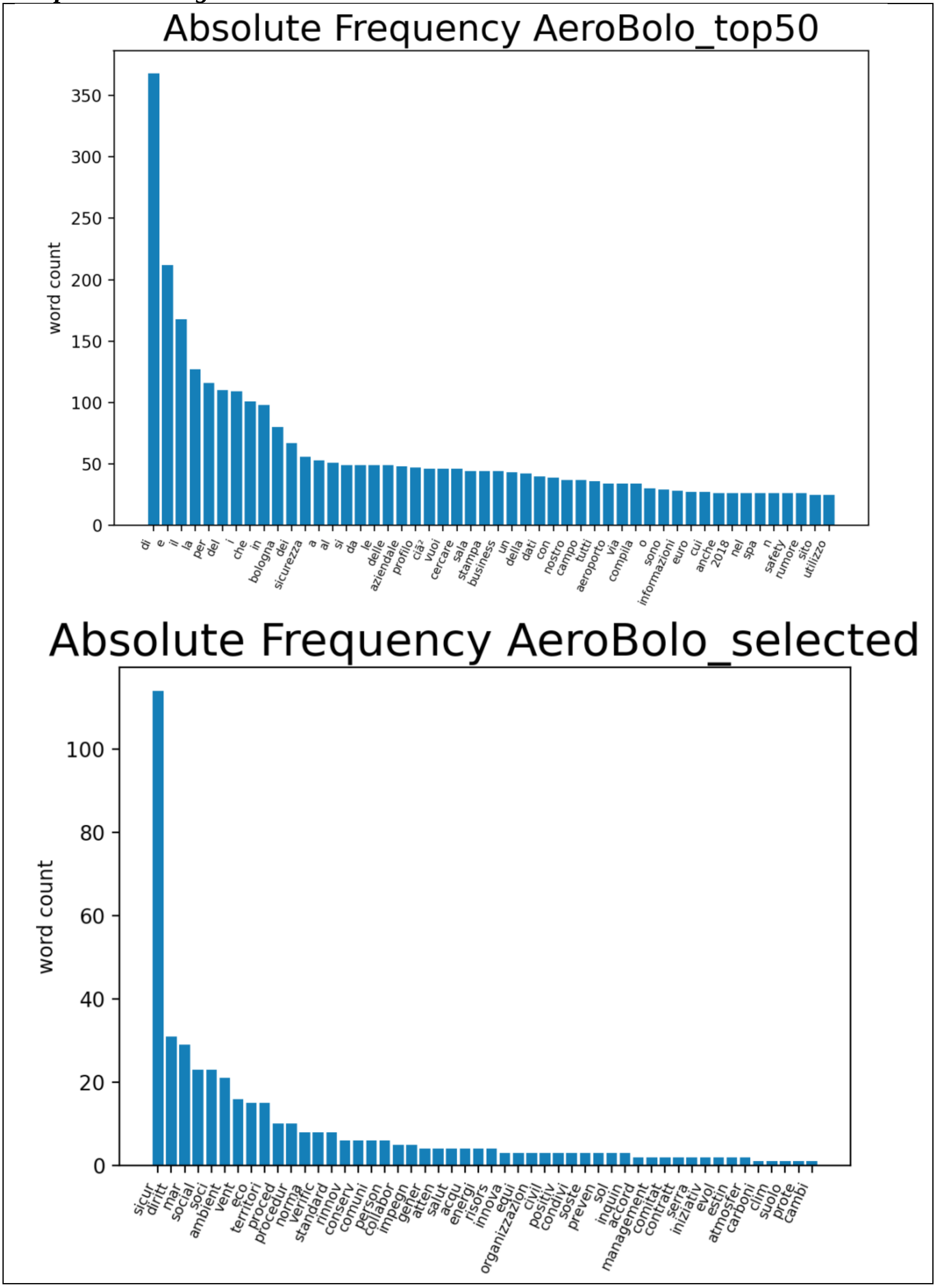

Absolute Frequency ACEA_selected
word count
0
20
40
60
80
100
120
140
sicur
territori
iniziativ
cultur
person
energi
ambient
vent
soci
comuni
gener
social
impegn
responsabil
innova
eco
sol
soste
mar
condivi
risors
natura
salut
collabor
evol
acqu
verific
inclus
atten
rinnov
cambi
positiv
management
cambiament
salvaguard
proced
procedur
traspare
standard
contratt
clim
verd
istituzion
amministr
preven
avanza
accord
unit
leader
prote

ACSM-AGAM

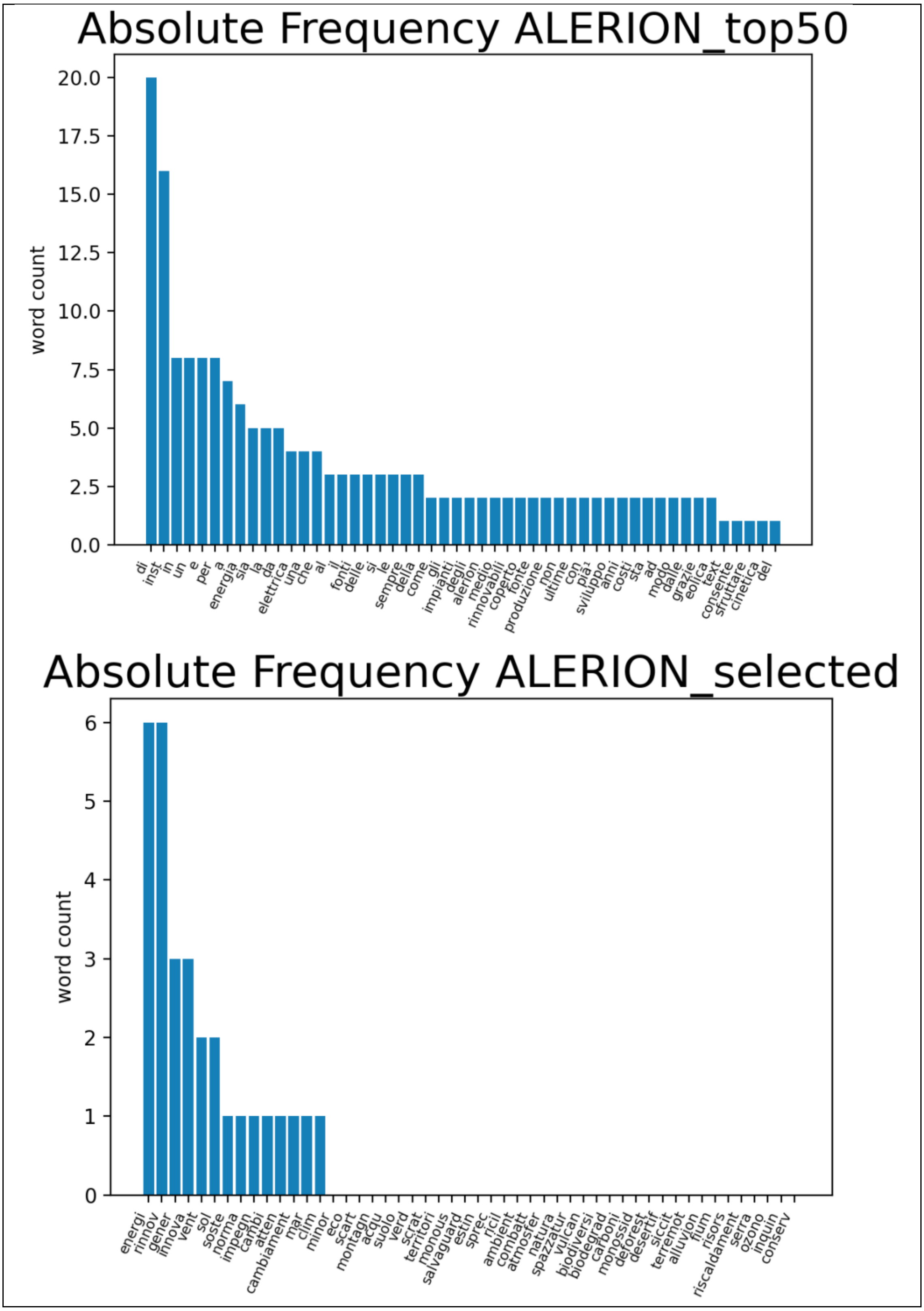

Absolute Frequency ACSM_top50
word count
0
2
4
6
8
10
12
14
di
e
delle
a
a2a
dei
reti
comune
energia
acsm
in
le
la
dal
agam
degli
smart
elettrica
linee
lario
sviluppo
ulteriore
nuovo
beneficio
ha
nel
da
un
del
con
progressivamente
cominciato
2018
luglio
primo
settore
protagonista
ruolo
consolidando
quotata
milano
borsa
alla
italia
nord
multiutility
principali
una
text
ad


Absolute Frequency ACSM_selected
word count
0.0
0.5
1.0
1.5
2.0
2.5
3.0
energi
ambient
gener
mar
risors
sol
innova
salvaguard
vent
acqu
suolo
verd
minor
territori
cambiament
monous
ricicl
estin
sprec
scart
atten
scrat
combatt
conserv
prote
evol
permess
montagn
spazzatur
eco
terremot
biodiversi
biodegrad
carboni
clim
monossid
deforest
desertif
siccit
alluvion
atmosfer
fium
natura
riscaldament
serra
rinnov
ozono
inquin
vulcan
preven

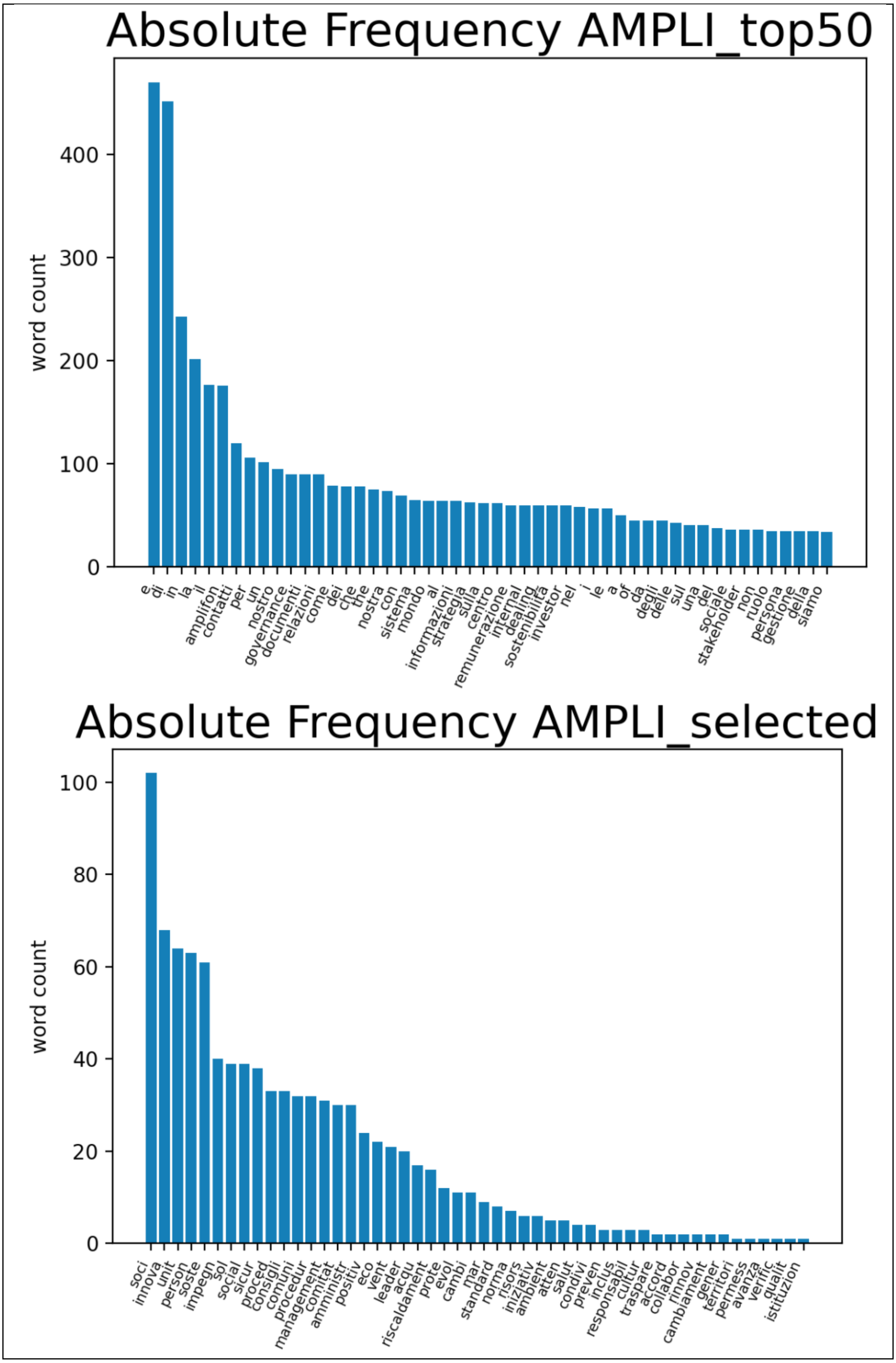
Aeroporto di Bologna



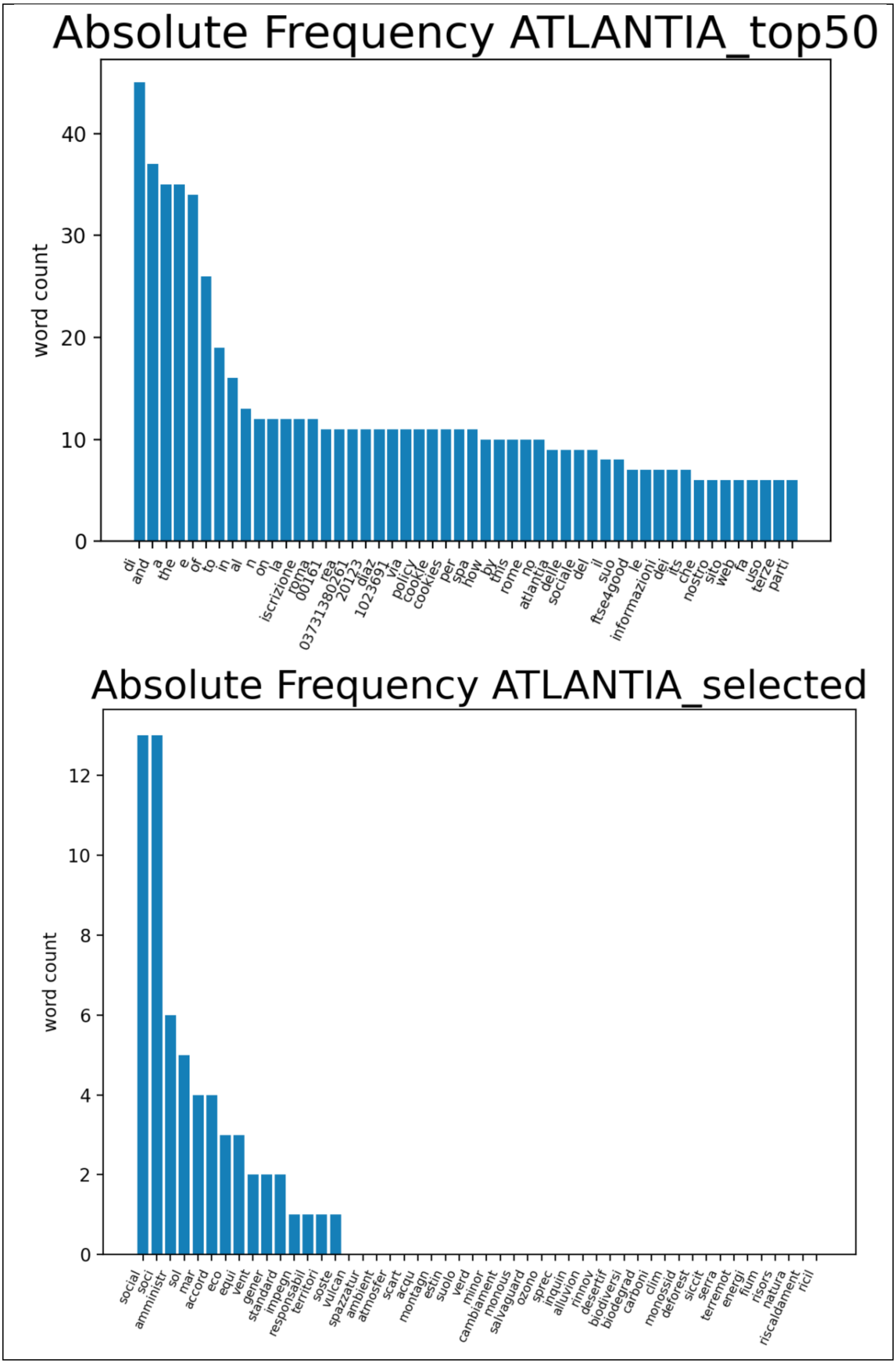
ALERION CLEAN ENERGY



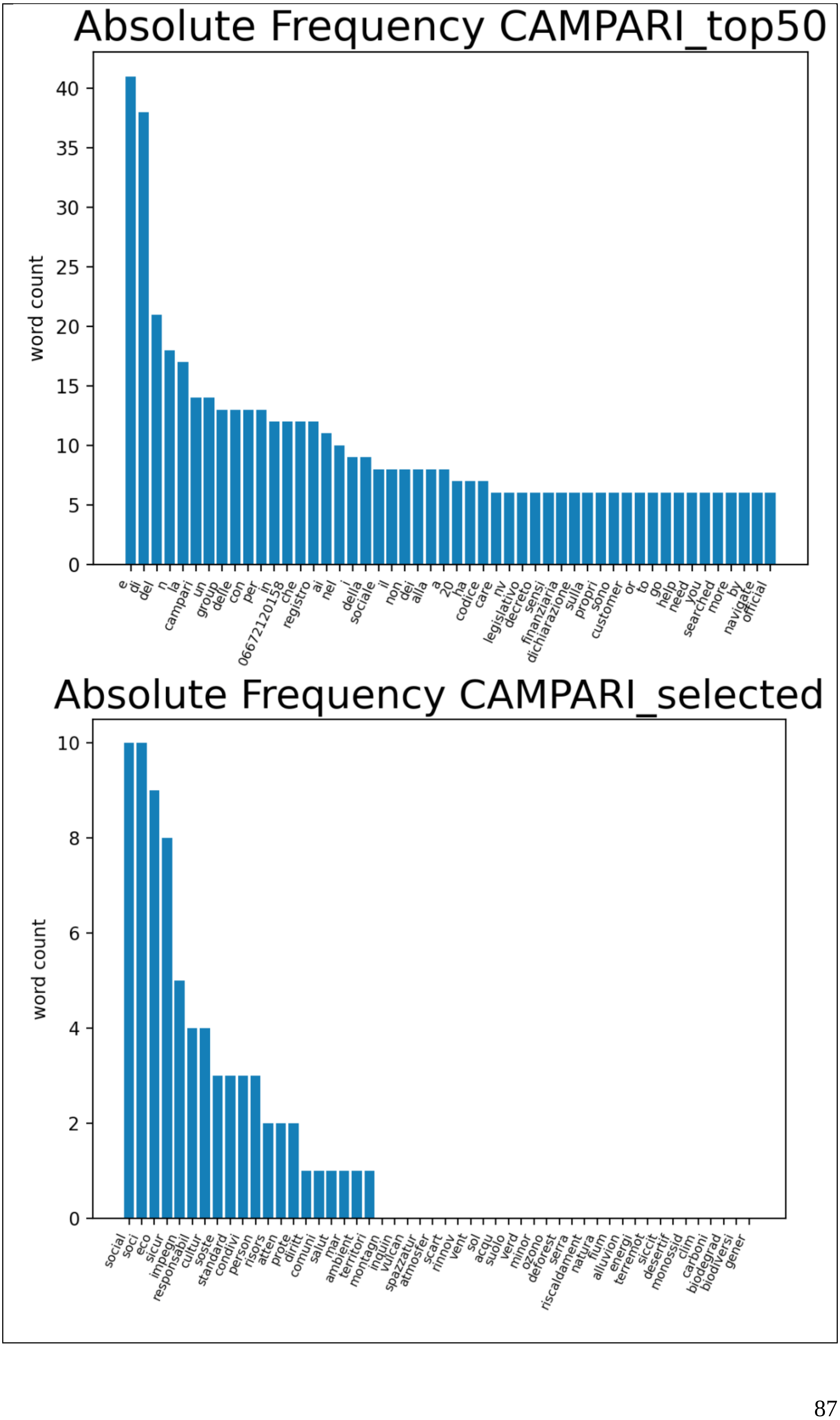
AMPLIFON


Absolute Frequency AMPLI_top50
word count
0
100
200
300
400
e
di
in
la
il
amplifon
contatti
per
un
nostro
governance
documenti
relazioni
come
dei
che
the
nostra
con
sistema
mondo
al
informazioni
strategia
sulla
centro
remunerazione
internal
dealing
sostenibilità
investor
nel
i
le
a
of
da
degli
delle
sul
una
del
sociale
stakeholder
non
ruolo
persona
gestione
della
siamo
Absolute Frequency AMPLI_selected
word count
0
20
40
60
80
100
soci
innova
unit
person
soste
impegn
sol
social
sicur
proced
consigli
comuni
procedur
management
comitat
amministr
positiv
eco
vent
leader
acqu
riscaldament
prote
evol
cambi
mar
standard
norma
risors
iniziativ
ambient
atten
salut
condivi
preven
inclus
responsabil
cultur
traspare
accord
collabor
rinnov
cambiament
gener
territori
permess
avanza
verific
qualit
istituzion

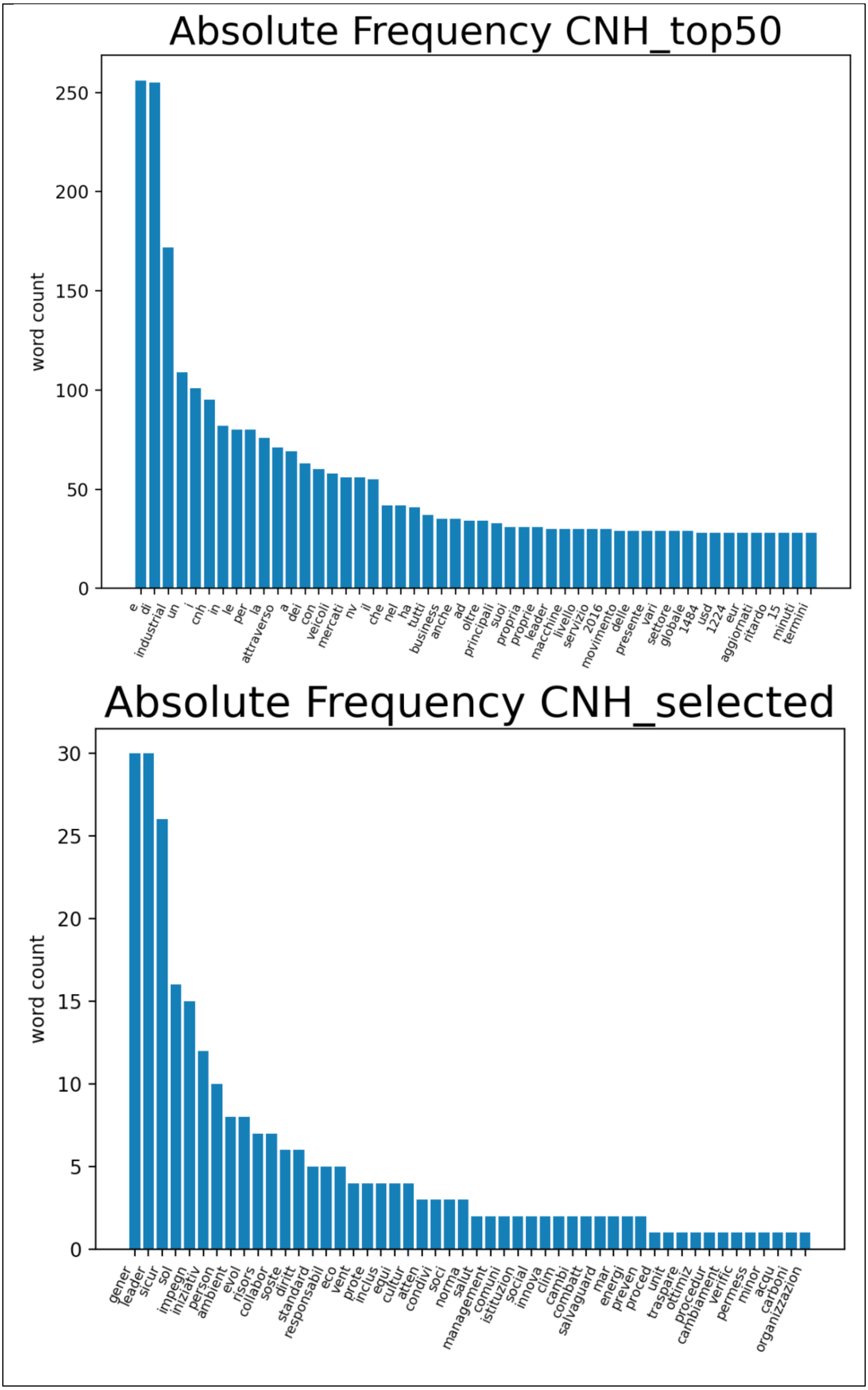
ATLANTIA


Absolute Frequency ATLANTIA_top50
word count
0
10
20
30
40
di
and
a
the
e
of
to
in
al
n
on
la
iscrizione
roma
00161
rea
03731380261
20123
diaz
1023691
via
policy
cookie
cookies
per
spa
how
by
this
rome
no
atlantia
delle
sociale
del
il
suo
ftse4good
le
informazioni
dei
its
che
nostro
sito
web
fa
uso
terze
parti
Absolute Frequency ATLANTIA_selected
word count
0
2
4
6
8
10
12
social
soci
amministr
sol
mar
accord
eco
equi
vent
gener
standard
impegn
responsabil
territori
soste
vulcan
spazzatur
ambient
atmosfer
scart
acqu
montagn
estin
suolo
verd
minor
cambiament
monous
salvaguard
ozono
sprec
inquin
alluvion
rinnov
desertif
biodiversi
biodegrad
carboni
clim
monossid
deforest
siccit
serra
terremot
energi
fium
risors
natura
riscaldament
ricil

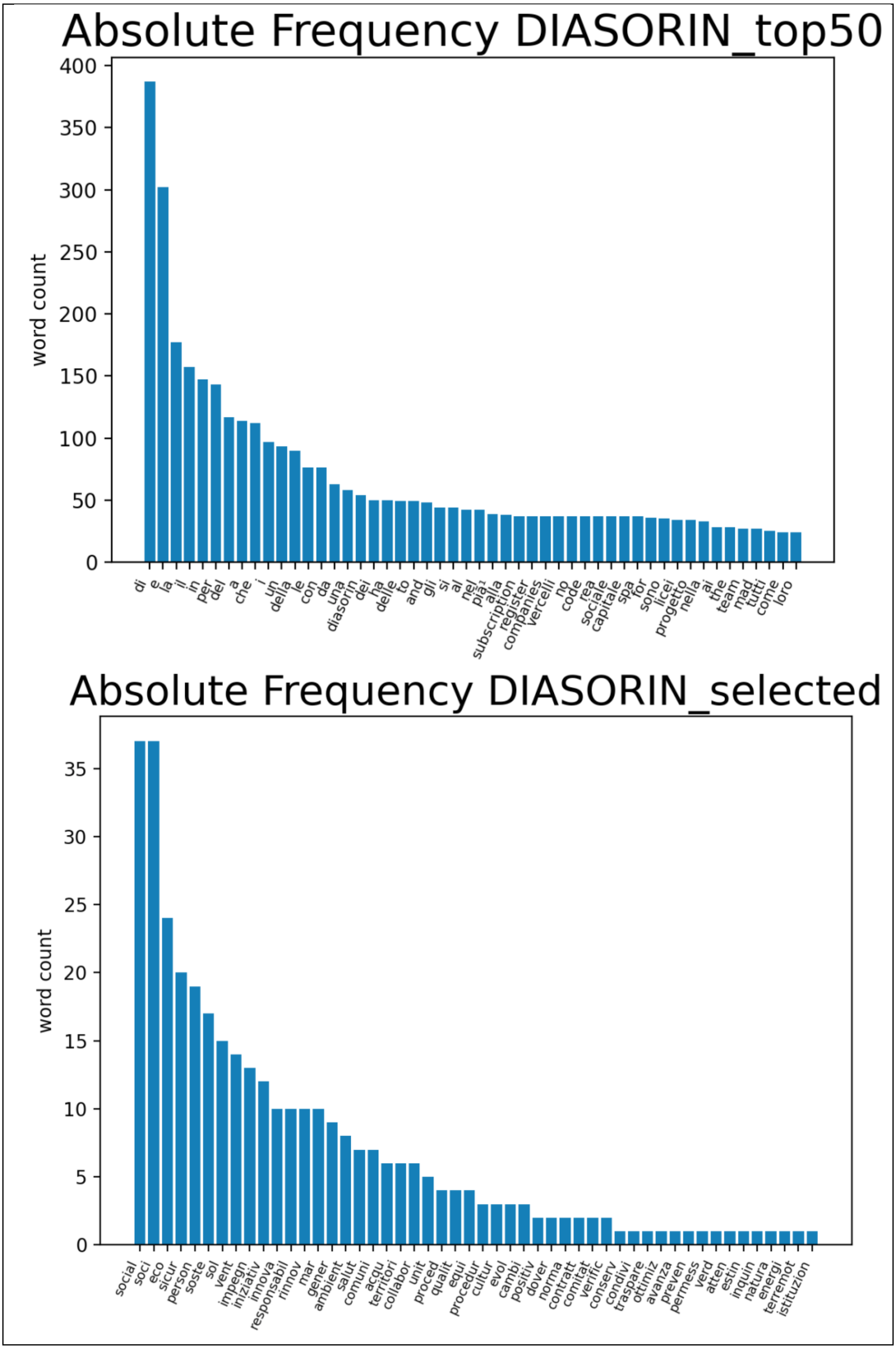
CAMPARI


Absolute Frequency CAMPARI_top50
word count
0
5
10
15
20
25
30
35
40
e
di
del
n
la
campari
un
group
delle
con
per
in
06672120158
che
registro
ai
nel
i
della
sociale
il
non
dei
alla
a
20
ha
codice
care
nv
legislativo
decreto
sensi
finanziaria
dichiarazione
sulla
propri
sono
customer
or
to
go
help
need
you
searched
more
by
navigate
official
Absolute Frequency CAMPARI_selected
word count
0
2
4
6
8
10
social
soci
eco
sicur
impegn
responsabil
cultur
soste
standard
condivi
person
risors
atten
prote
diritt
comuni
salut
mar
ambient
territori
montagn
inquin
vulcan
spazzatur
atmosfer
scart
rinnov
vent
sol
acqu
suolo
verd
minor
ozono
deforest
serra
riscaldament
natura
fium
alluvion
energi
terremot
siccit
desertif
monossid
clim
carboni
biodegrad
biodiversi
gener

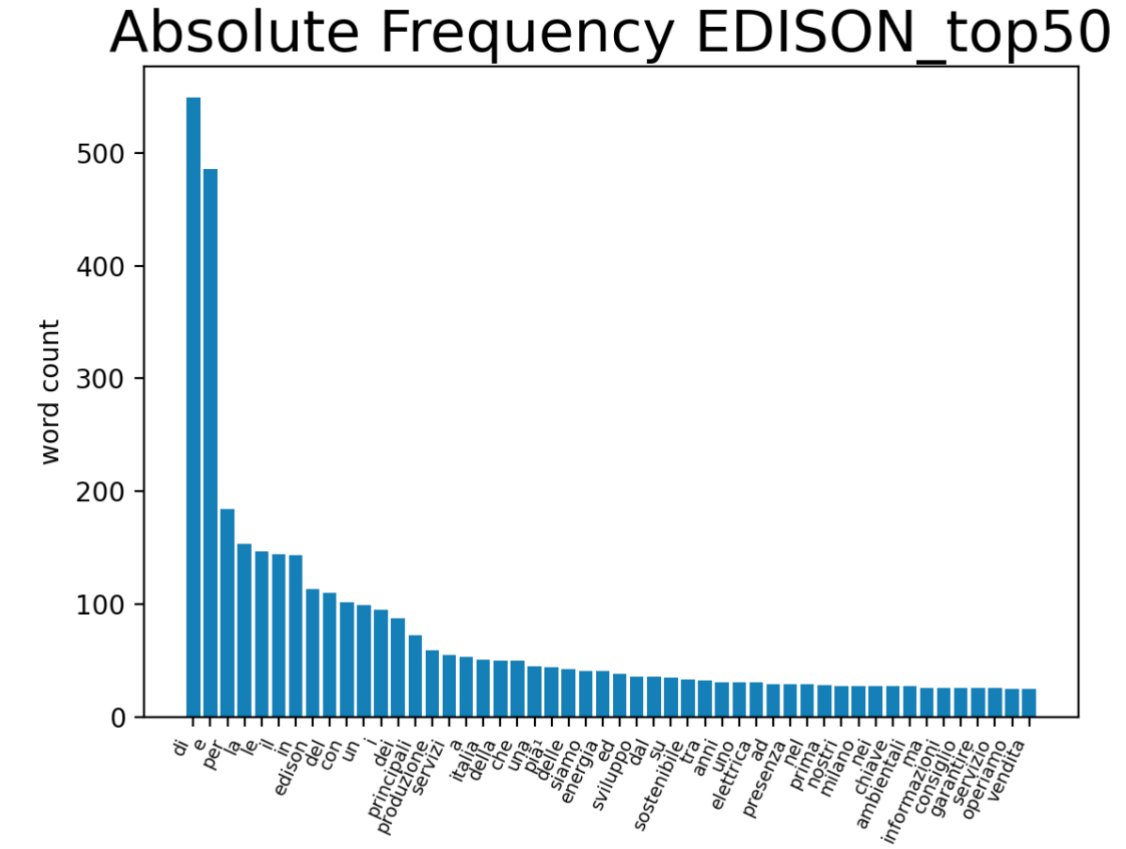
CNH

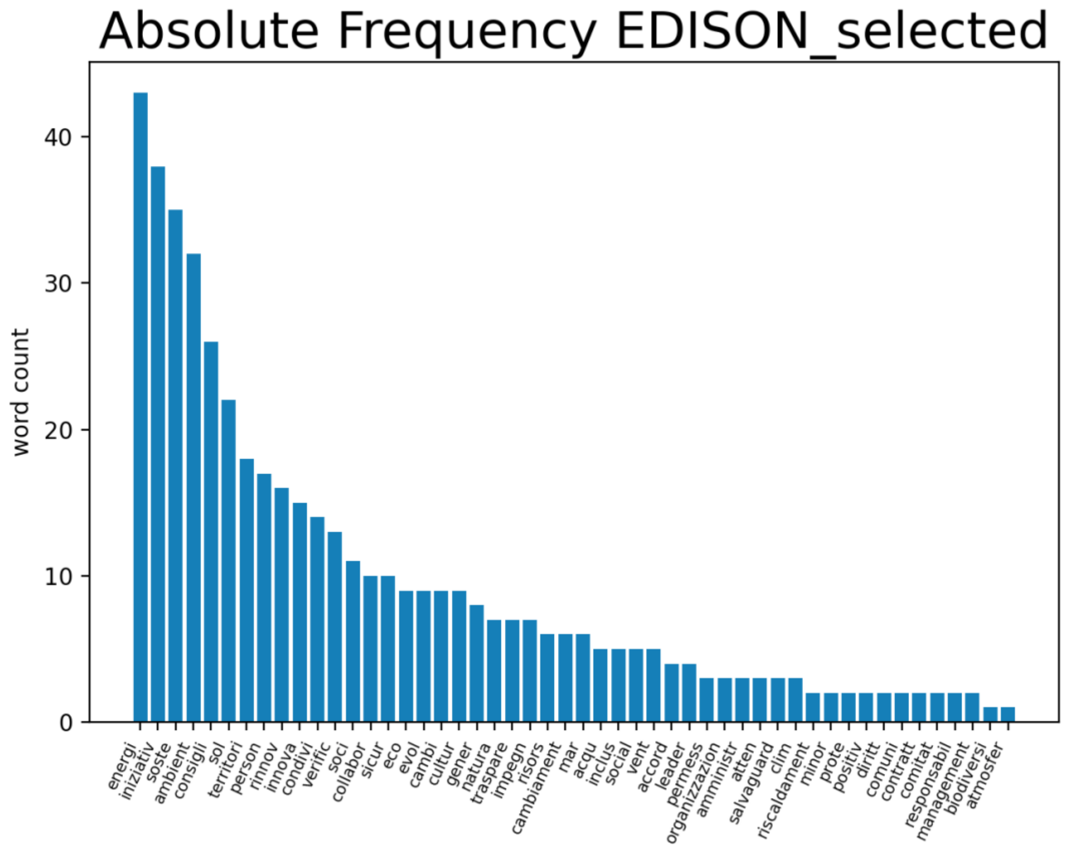

Absolute Frequency CNH_top50
word count
0
50
100
150
200
250
e
di
industrial
un
i
cnh
in
le
per
la
attraverso
a
dei
con
veicoli
mercati
nv
il
che
nel
ha
tutti
business
anche
ad
oltre
principali
suoi
propria
proprie
leader
macchine
livello
servizio
2016
movimento
delle
presente
vari
settore
globale
1484
usd
1224
eur
aggiornati
ritardo
15
minuti
termini
Absolute Frequency CNH_selected
word count
0
5
10
15
20
25
30
gener
leader
sicur
sol
impegn
iniziativ
person
ambient
evol
risors
collabor
soste
diritt
standard
responsabil
eco
vent
prote
inclus
equi
cultur
atten
condivi
soci
norma
salut
management
comuni
istituzion
social
innova
clim
cambi
combatt
salvaguard
mar
energi
preven
proced
unit
traspare
ottimiz
procedur
cambiament
verific
permess
minor
acqu
carboni
organizzazion

DIASORIN

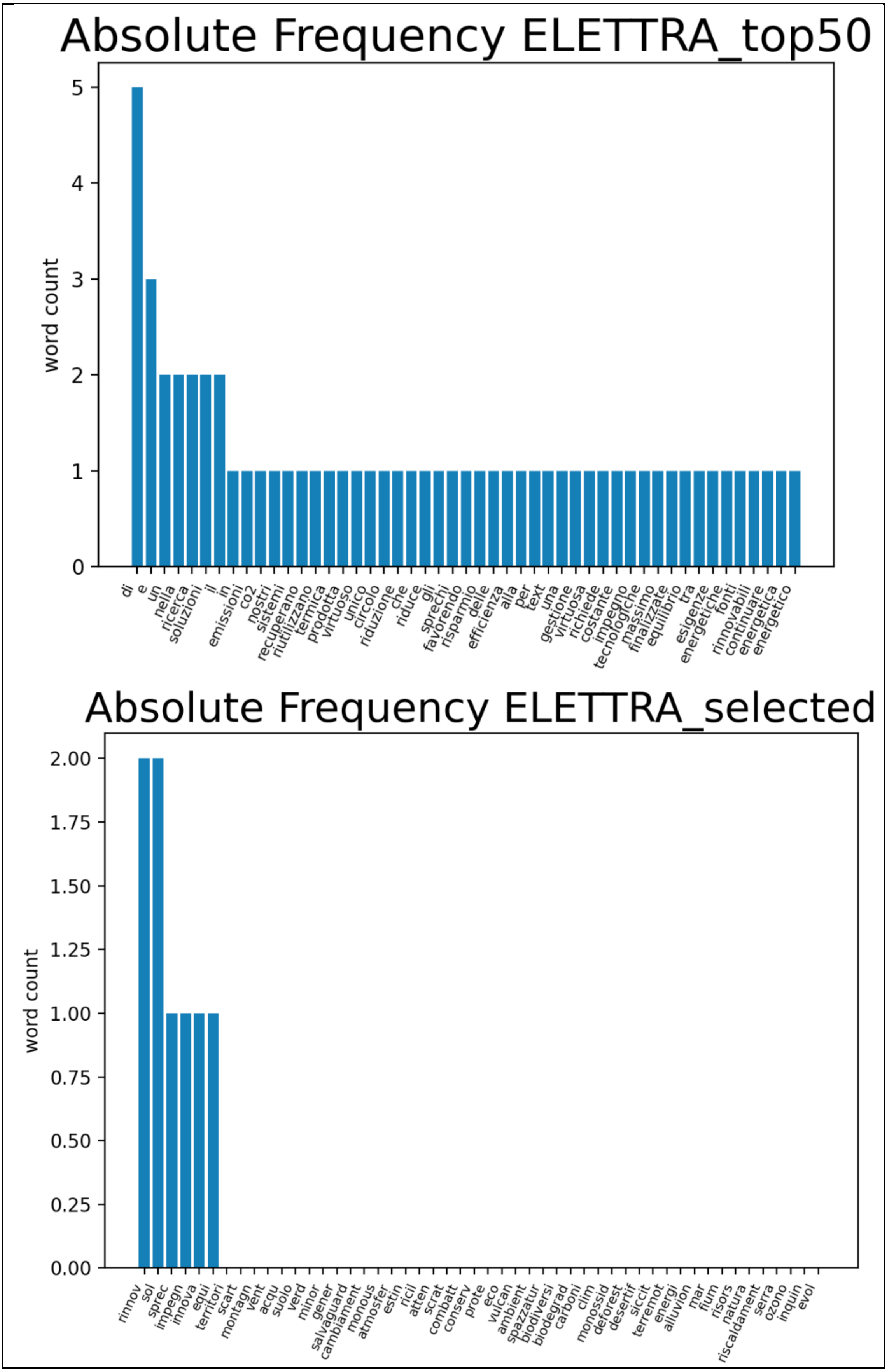

Absolute Frequency DIASORIN_top50
word count
0
50
100
150
200
250
300
350
400
di
e
la
il
in
per
del
a
che
i
un
della
le
con
da
una
diasorin
dei
ha
delle
to
and
gli
si
al
nel
pià¹
alla
subscription
register
companies
vercelli
no
code
rea
sociale
capitale
spa
for
sono
licei
progetto
nella
ai
the
team
mad
tutti
come
loro
Absolute Frequency DIASORIN_selected
word count
0
5
10
15
20
25
30
35
social
soci
eco
sicur
person
soste
sol
vent
impegn
iniziativ
innova
responsabil
rinnov
mar
gener
ambient
salut
comuni
acqu
territori
collabor
unit
proced
qualit
equi
procedur
cultur
evol
cambi
positiv
dover
norma
contratt
comitat
verific
conserv
condivi
traspare
ottimiz
avanza
preven
permess
verd
atten
estin
inquin
natura
energi
terremot
istituzion

EDISON

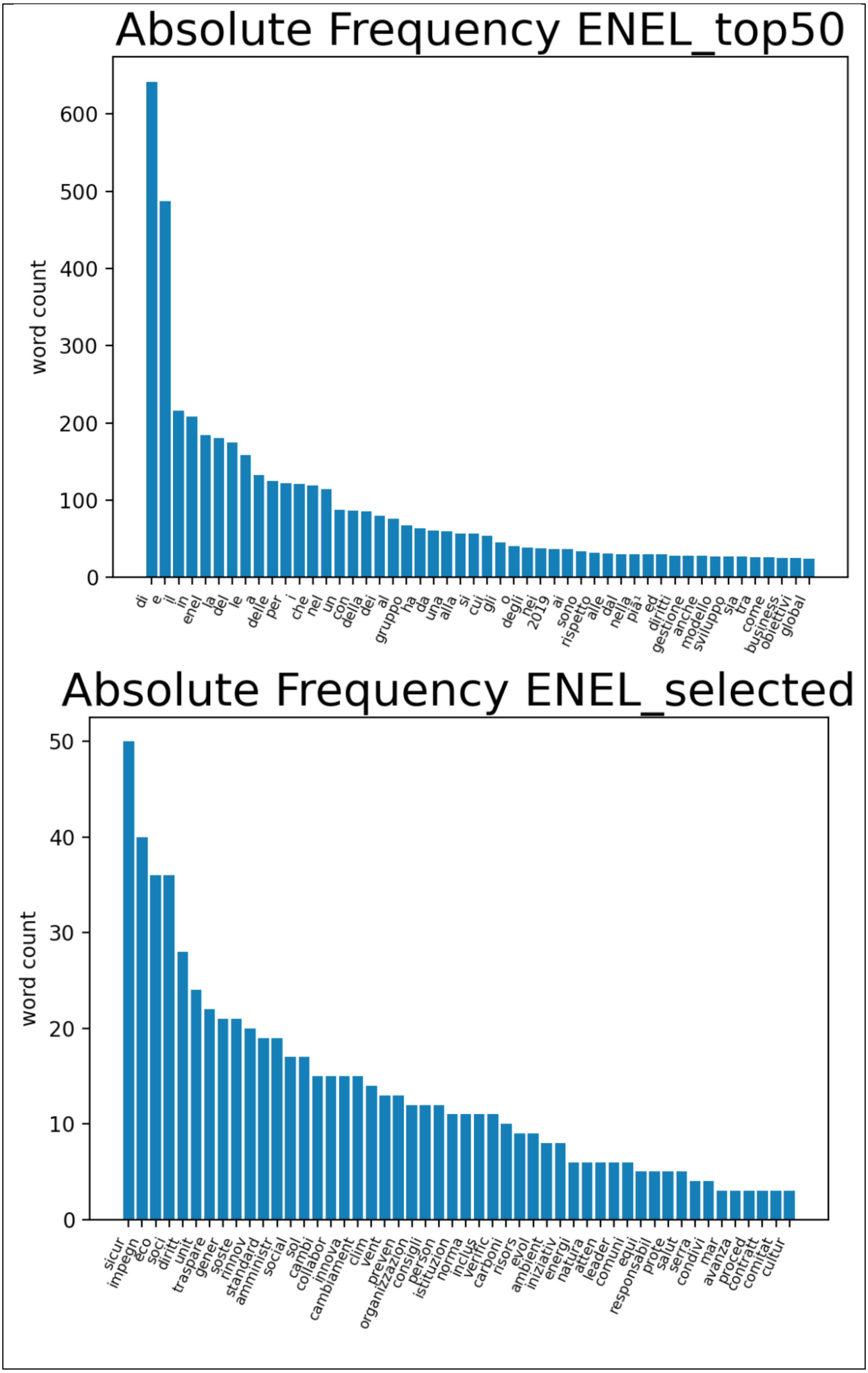

Absolute Frequency EDISON_top50
word count
0
100
200
300
400
500
di
e
per
la
le
il
in
edison
del
con
un
i
dei
principali
produzione
servizi
a
italia
della
che
una
più
delle
siamo
energia
ed
sviluppo
dal
su
sostenibile
tra
anni
uno
elettrica
ad
presenza
nel
prima
nostri
milano
nei
chiave
ambientali
ma
informazioni
consiglio
garantire
servizio
operiamo
vendita


Absolute Frequency EDISON_selected
word count
0
10
20
30
40
energi
iniziativ
soste
ambient
consigli
sol
territori
person
rinnov
innova
condivi
verific
soci
collabor
sicur
eco
evol
cambi
cultur
gener
natura
traspare
impegn
risors
cambiament
mar
acqu
inclus
social
vent
accord
leader
permess
organizzazion
amministr
atten
salvaguard
clim
riscaldament
minor
prote
positiv
diritt
comuni
contratt
comitat
responsabil
management
biodiversi
atmosfer

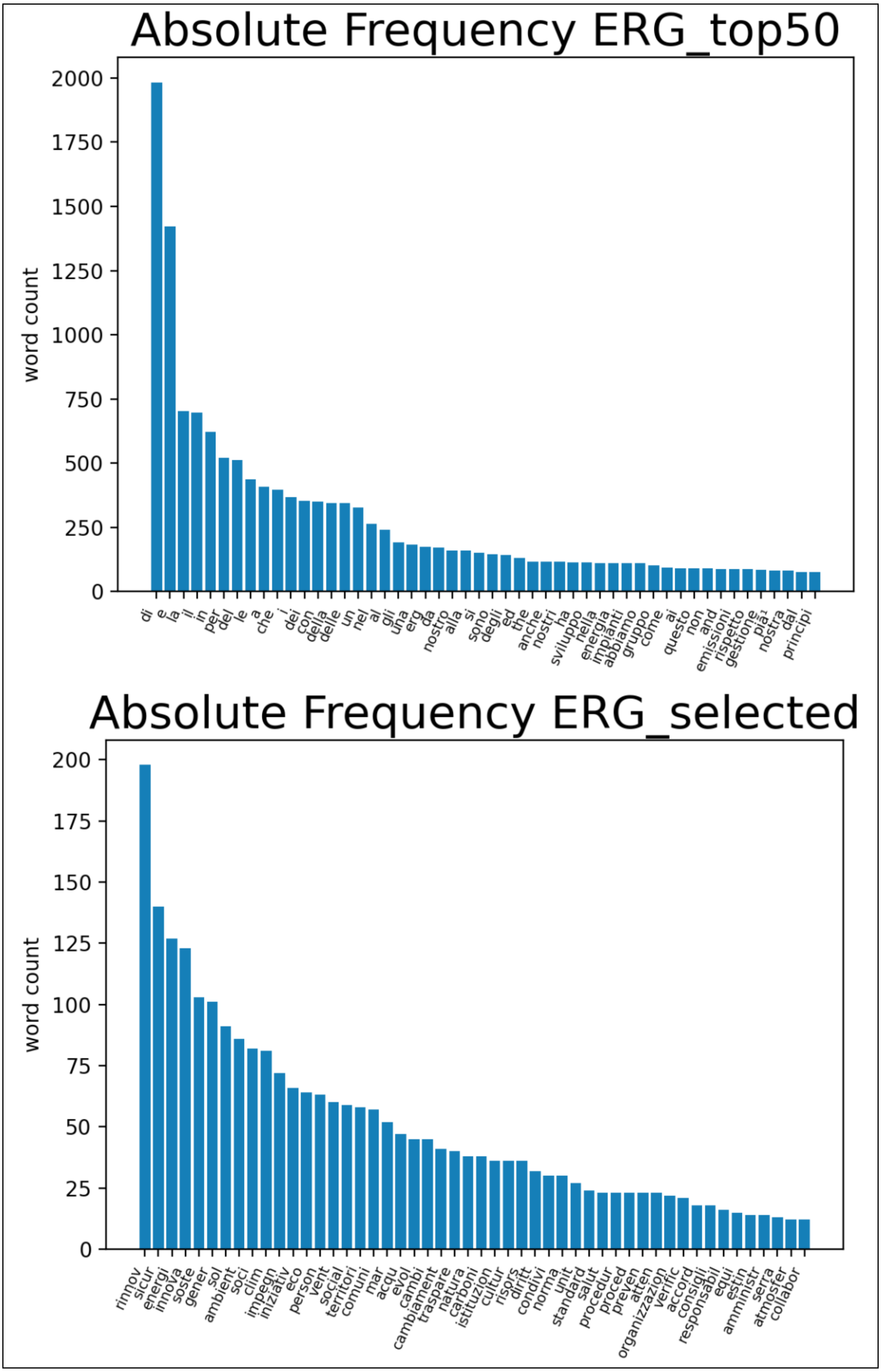
ELETTRA ENERGIA



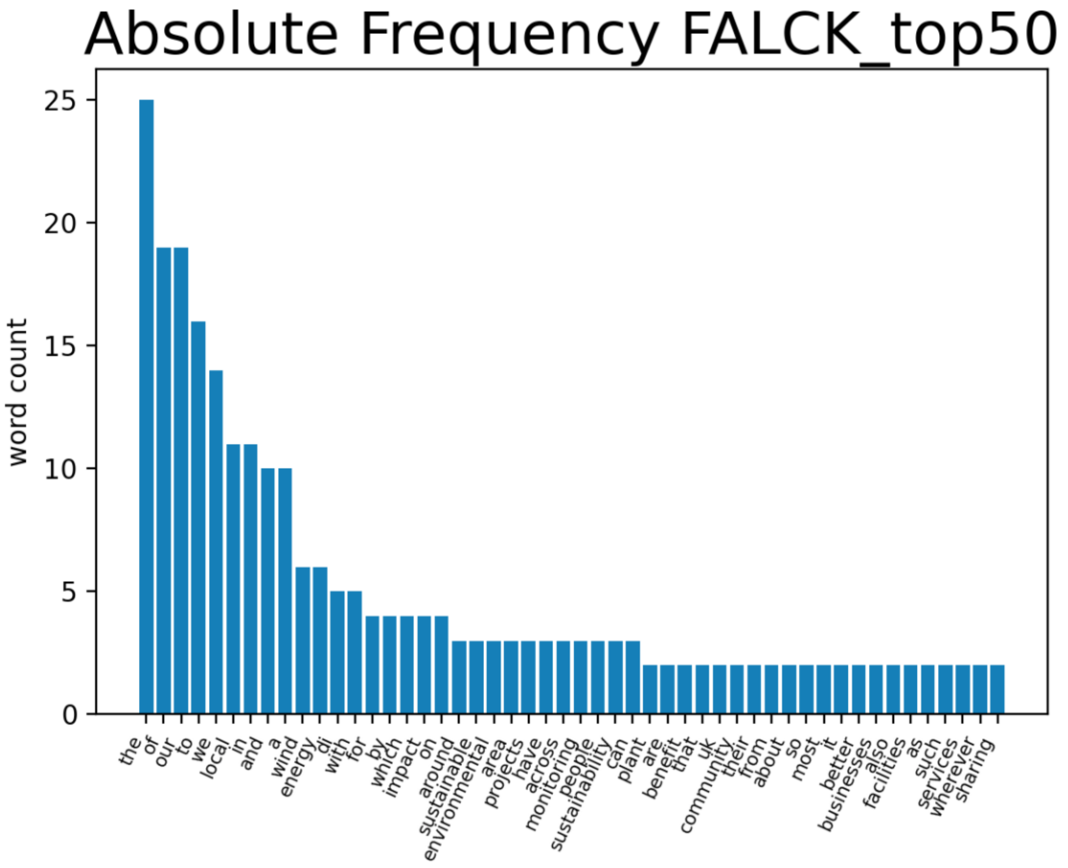
ENEL

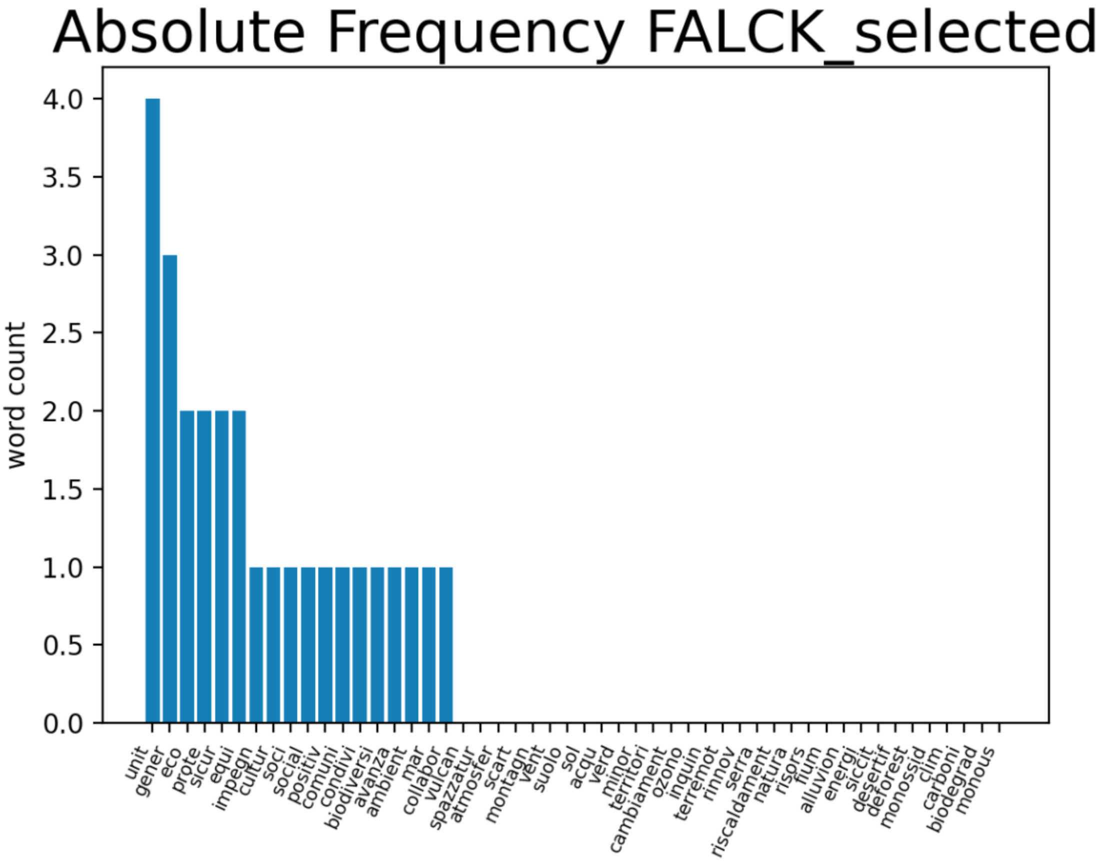

Absolute Frequency ENEL_top50
word count
0
100
200
300
400
500
600
di
e
il
in
enel
la
del
le
a
delle
per
i
che
nel
un
con
della
dei
al
gruppo
ha
da
una
alla
si
cui
gli
o
degli
nei
2019
ai
sono
rispetto
alle
dal
nella
piã¹
ed
diritti
gestione
anche
modello
sviluppo
sia
tra
come
business
obiettivi
global
Absolute Frequency ENEL_selected
word count
0
10
20
30
40
50
sicur
impegn
eco
soci
diritt
unit
traspare
gener
soste
rinnov
standard
amministr
social
sol
cambi
collabor
innova
cambiament
clim
vent
preven
organizzazion
consigli
person
istituzion
norma
inclus
verific
carboni
risors
evol
ambient
iniziativ
energi
natura
atten
leader
comuni
equi
responsabil
prote
salut
serra
condivi
mar
avanza
proced
contratt
comitat
cultur

ERG

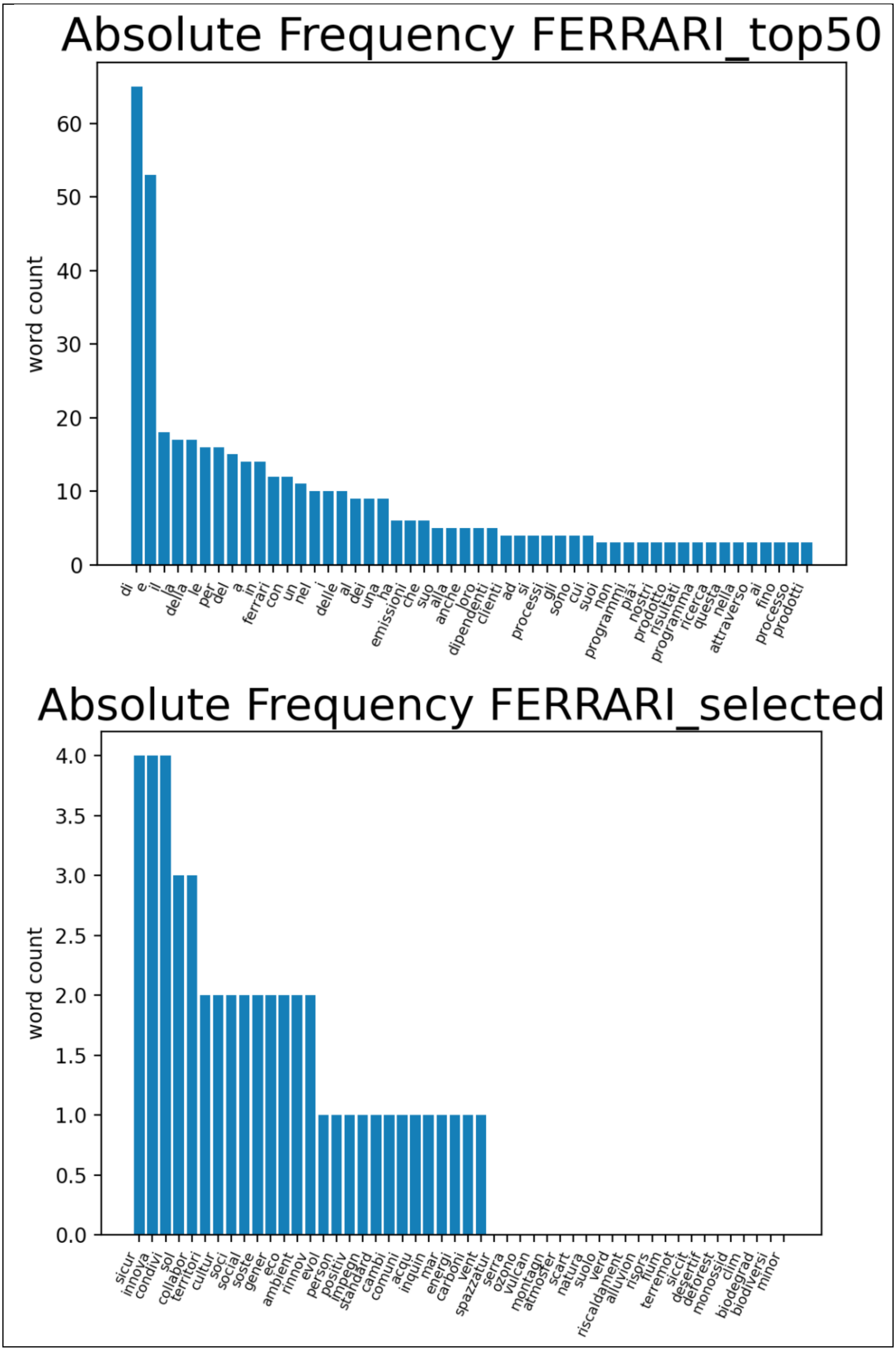

Absolute Frequency ERG_top50
word count
2000
1750
1500
1250
1000
750
500
250
0
di
e
la
il
in
per
del
le
a
che
i
dei
con
della
delle
un
nel
al
gli
una
erg
da
nostro
alla
si
sono
degli
ed
the
anche
nostri
ha
sviluppo
nella
energia
impianti
abbiamo
gruppo
come
ai
questo
non
and
emissioni
rispetto
gestione
piã¹
nostra
dal
principi
Absolute Frequency ERG_selected
word count
200
175
150
125
100
75
50
25
0
rinnov
sicur
energi
innova
soste
gener
sol
ambient
soci
clim
impegn
iniziativ
eco
person
vent
social
territori
comuni
mar
acqu
evol
cambi
cambiament
traspare
natura
carboni
istituzion
cultur
risors
diritt
condivi
norma
unit
standard
salut
procedur
proced
preven
atten
organizzazion
verific
accord
consigli
responsabil
equi
estin
amministr
serra
atmosfer
collabor

FALCK RENEWABLES

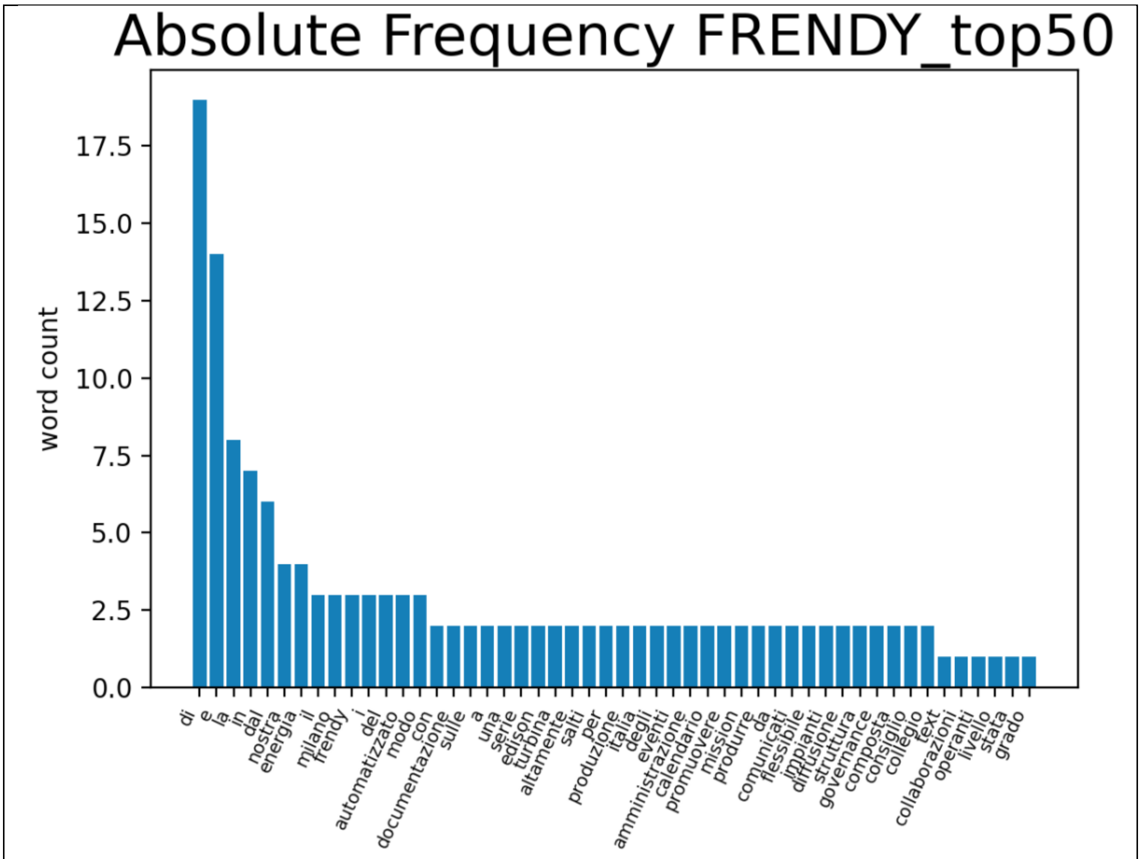


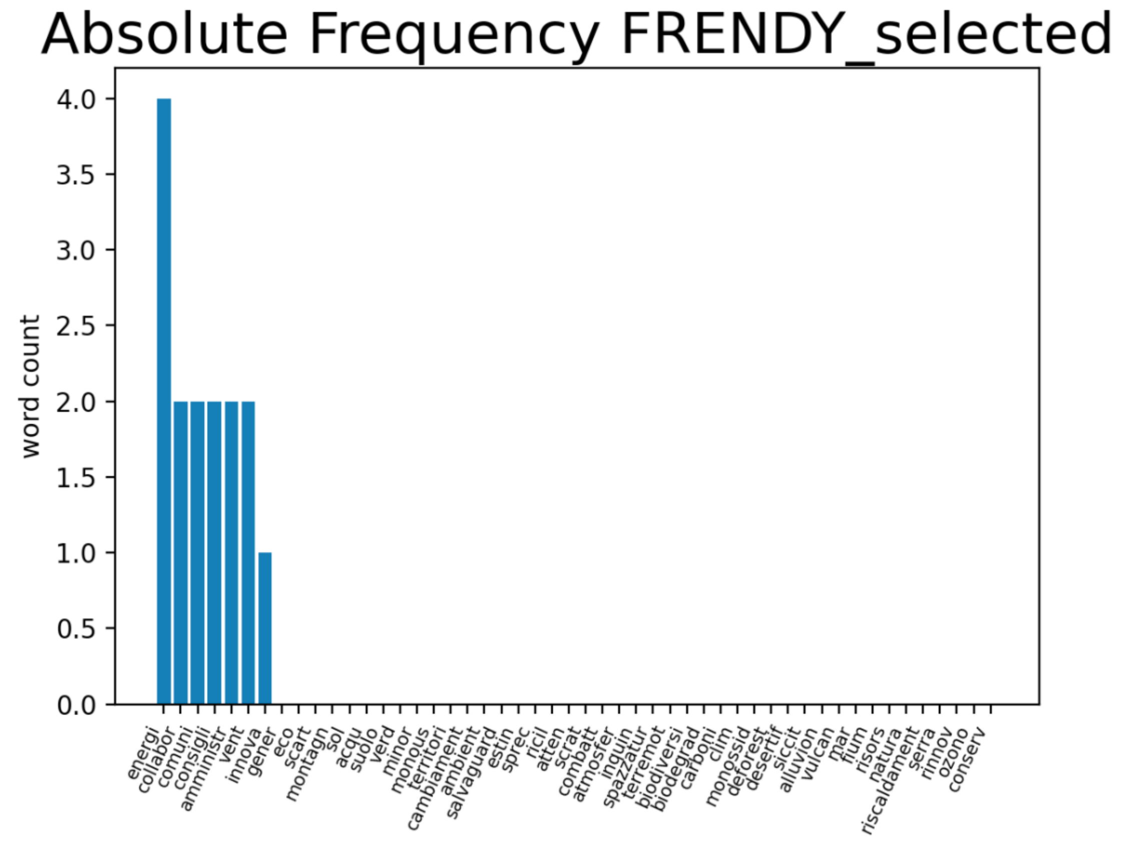


FERRARI

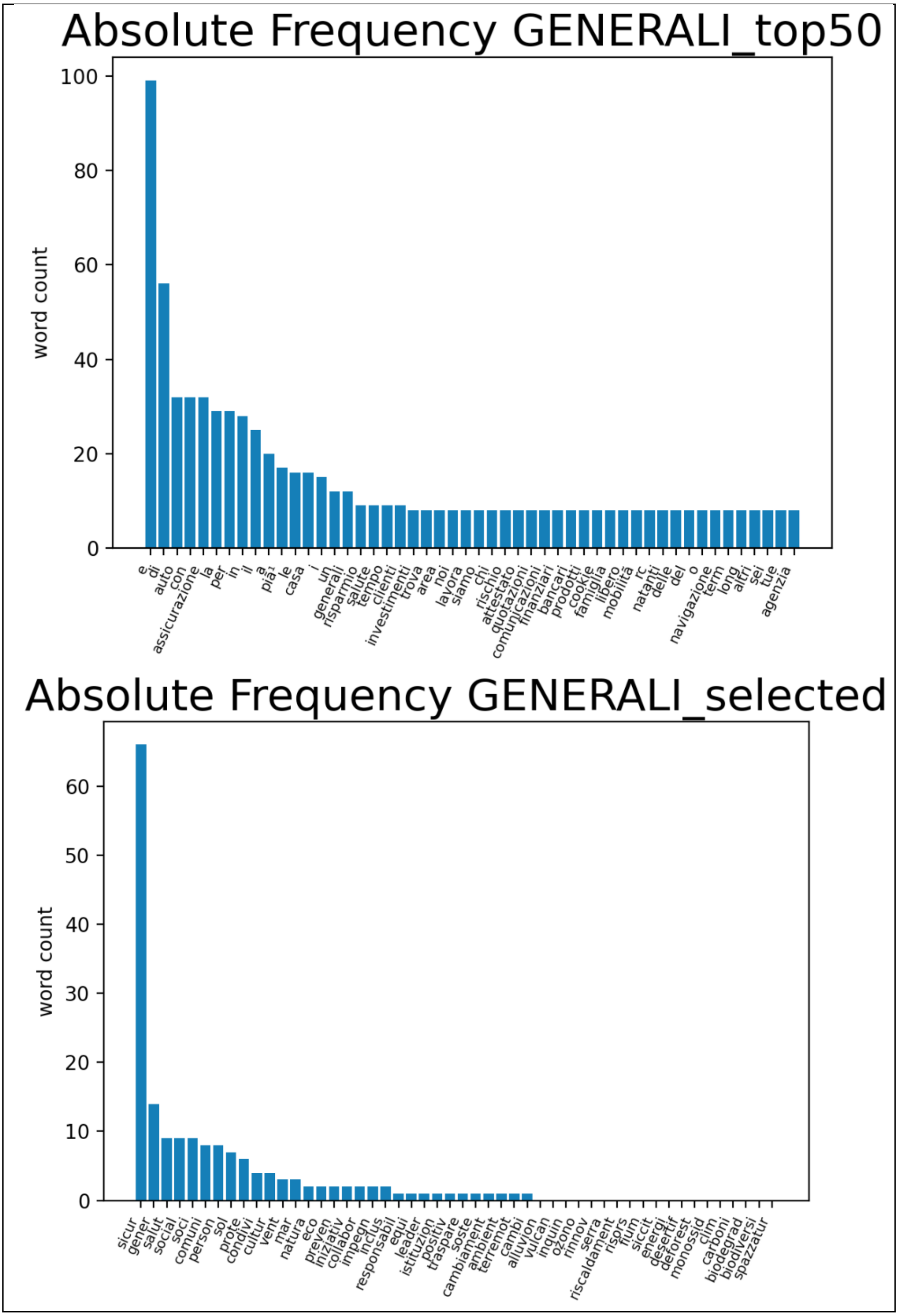

Absolute Frequency FERRARI_top50
word count
0
10
20
30
40
50
60
di
e
il
la
della
le
per
del
a
in
ferrari
con
un
nel
i
delle
al
dei
una
ha
emissioni
che
suo
alla
anche
loro
dipendenti
clienti
ad
si
processi
gli
sono
cui
suoi
non
programmi
piã¹
nostri
prodotto
risultati
programma
ricerca
questa
nella
attraverso
ai
fino
processo
prodotti
Absolute Frequency FERRARI_selected
word count
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
sicur
innova
condivi
sol
collabor
territori
cultur
soci
social
soste
gener
eco
ambient
rinnov
evol
person
positiv
impegn
standard
cambi
comuni
acqu
inquin
mar
energi
carboni
vent
spazzatur
serra
ozono
vulcan
montagn
atmosfer
scart
natura
suolo
verd
riscaldament
alluvion
risors
fium
terremot
siccit
desertif
deforest
monossid
clim
biodegrad
biodiversi
minor

FRENDY ENERGY

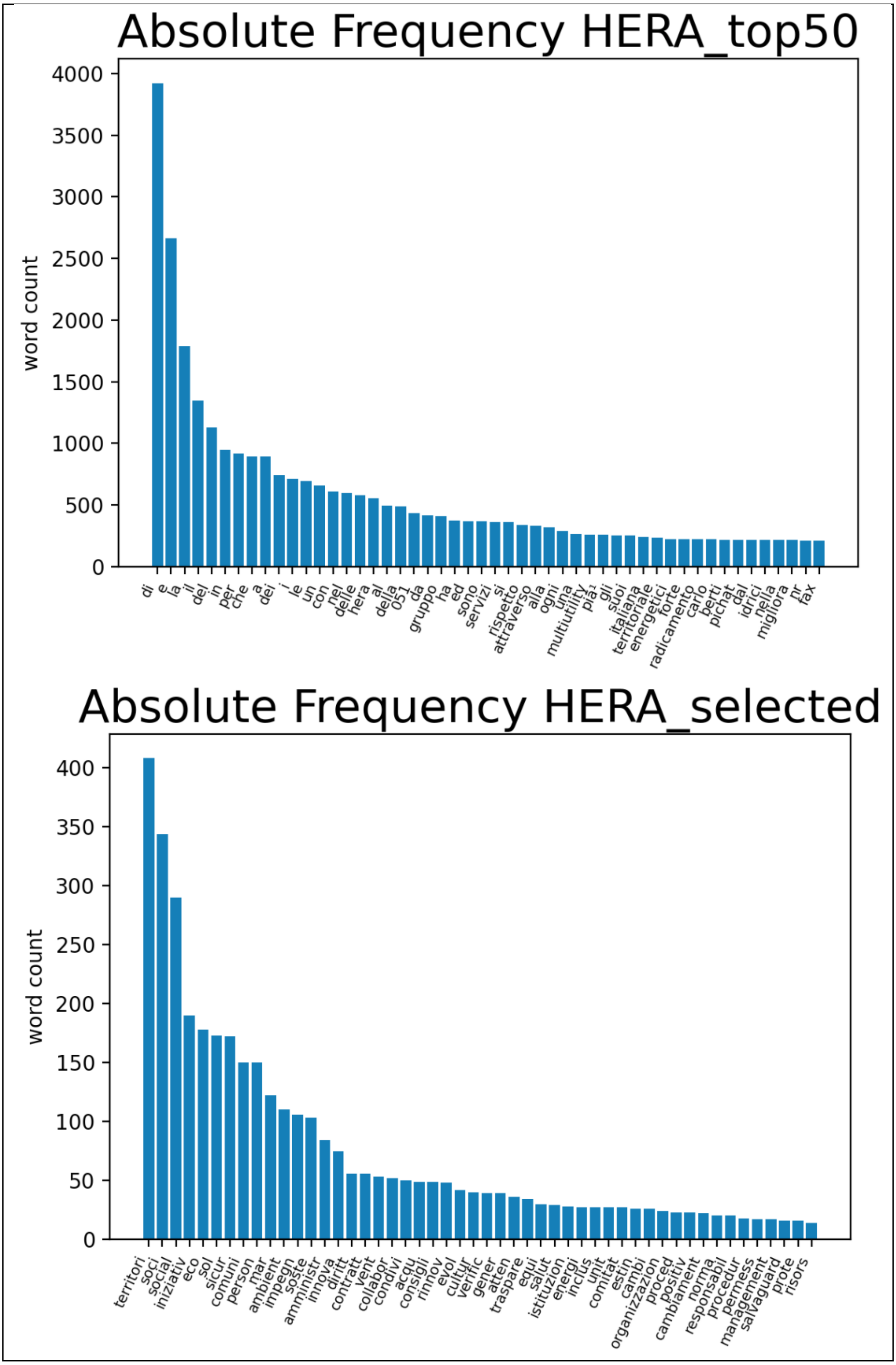




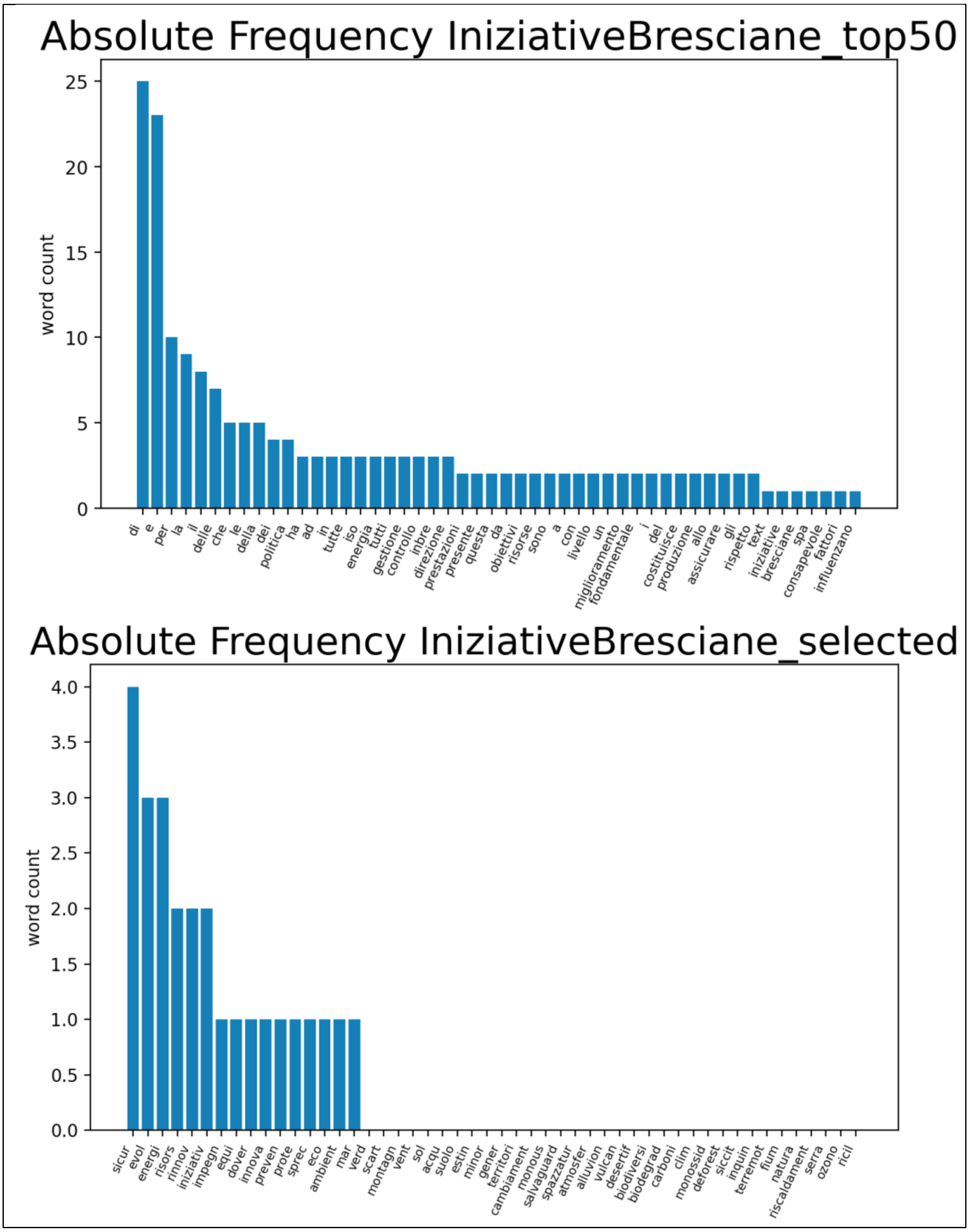
GENERALI


Absolute Frequency GENERALI_top50
word count
100
80
60
40
20
0
e
di
auto
con
assicurazione
la
per
in
il
a
piã¹
le
casa
i
un
generali
risparmio
salute
tempo
clienti
investimenti
trova
area
noi
lavora
siamo
chi
rischio
attestato
quotazioni
comunicazioni
finanziari
bancari
prodotti
cookie
famiglia
libero
mobilità
rc
natanti
delle
del
o
navigazione
term
long
altri
sei
tue
agenzia
Absolute Frequency GENERALI_selected
word count
60
50
40
30
20
10
0
sicur
gener
salut
social
soci
comuni
person
sol
prote
condivi
cultur
vent
mar
natura
eco
preven
iniziativ
collabor
impegn
inclus
responsabil
equi
leader
istituzion
positiv
traspare
soste
cambiament
ambient
terremot
cambi
alluvion
vulcan
inquin
ozono
rinnov
serra
riscaldament
risors
fium
siccit
energi
desertif
deforest
monossid
clim
carboni
biodegrad
biodiversi
spazzatur

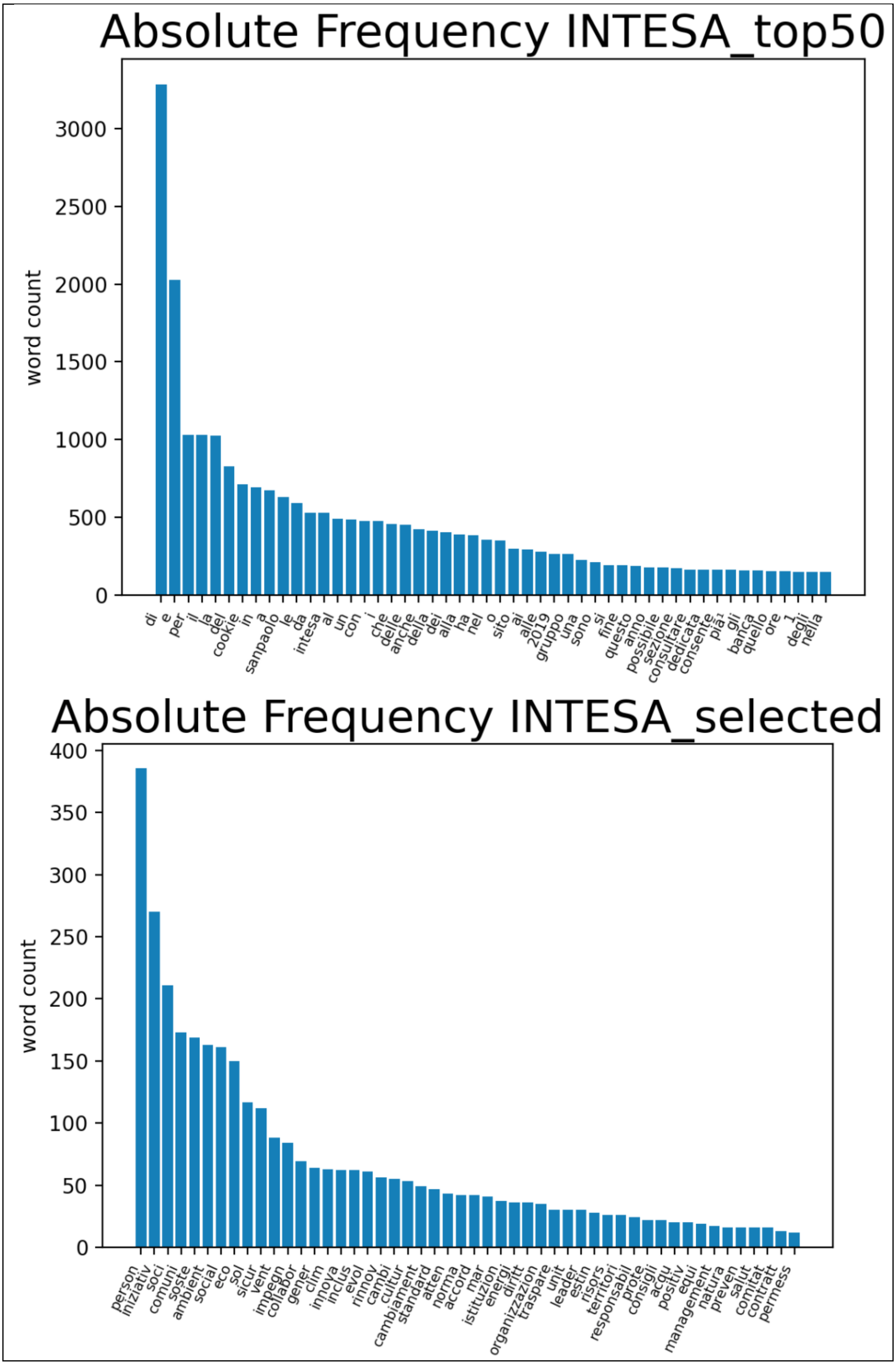
HERA


Absolute Frequency HERA_top50
word count
0
500
1000
1500
2000
2500
3000
3500
4000
di
e
la
il
del
in
per
che
a
dei
i
le
un
con
nel
delle
hera
al
della
051
da
gruppo
ha
ed
sono
servizi
si
rispetto
attraverso
alla
ogni
una
multiutility
piã¹
gli
suoi
italiana
territoriale
energetici
forte
radicamento
carlo
berti
pichat
dal
idrici
nella
migliora
nr
fax
Absolute Frequency HERA_selected
word count
0
50
100
150
200
250
300
350
400
territori
soci
social
iniziativ
eco
sol
sicur
comuni
person
mar
ambient
impegn
soste
amministr
innova
diritt
contratt
vent
collabor
condivi
acqu
consigli
rinnov
evol
cultur
verific
gener
atten
traspare
equi
salut
istituzion
energi
inclus
unit
comitat
estin
cambi
organizzazion
proced
positiv
cambiament
norma
responsabil
procedur
permess
management
salvaguard
prote
risors

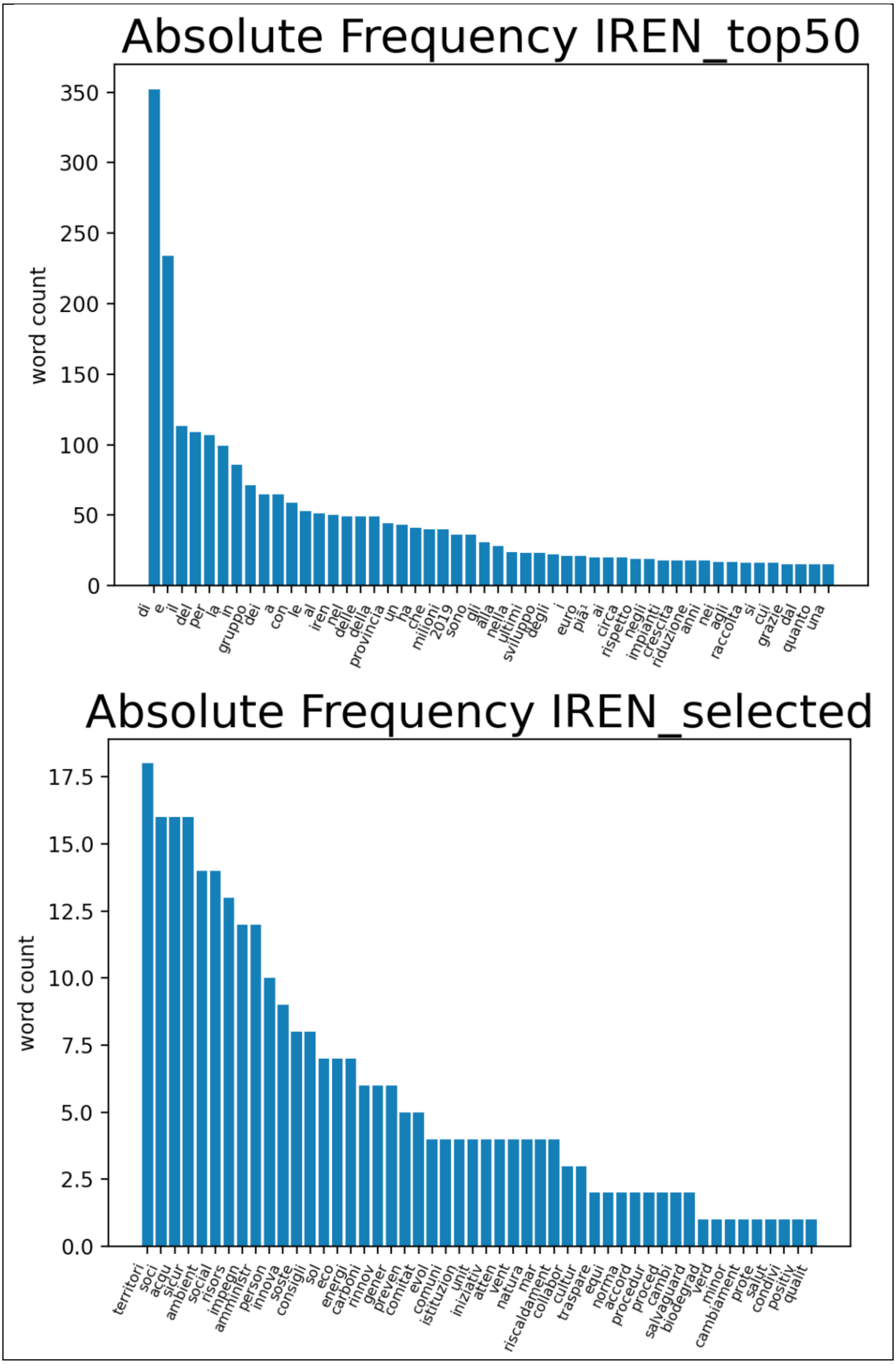
INIZIATIVE BRESCIANE



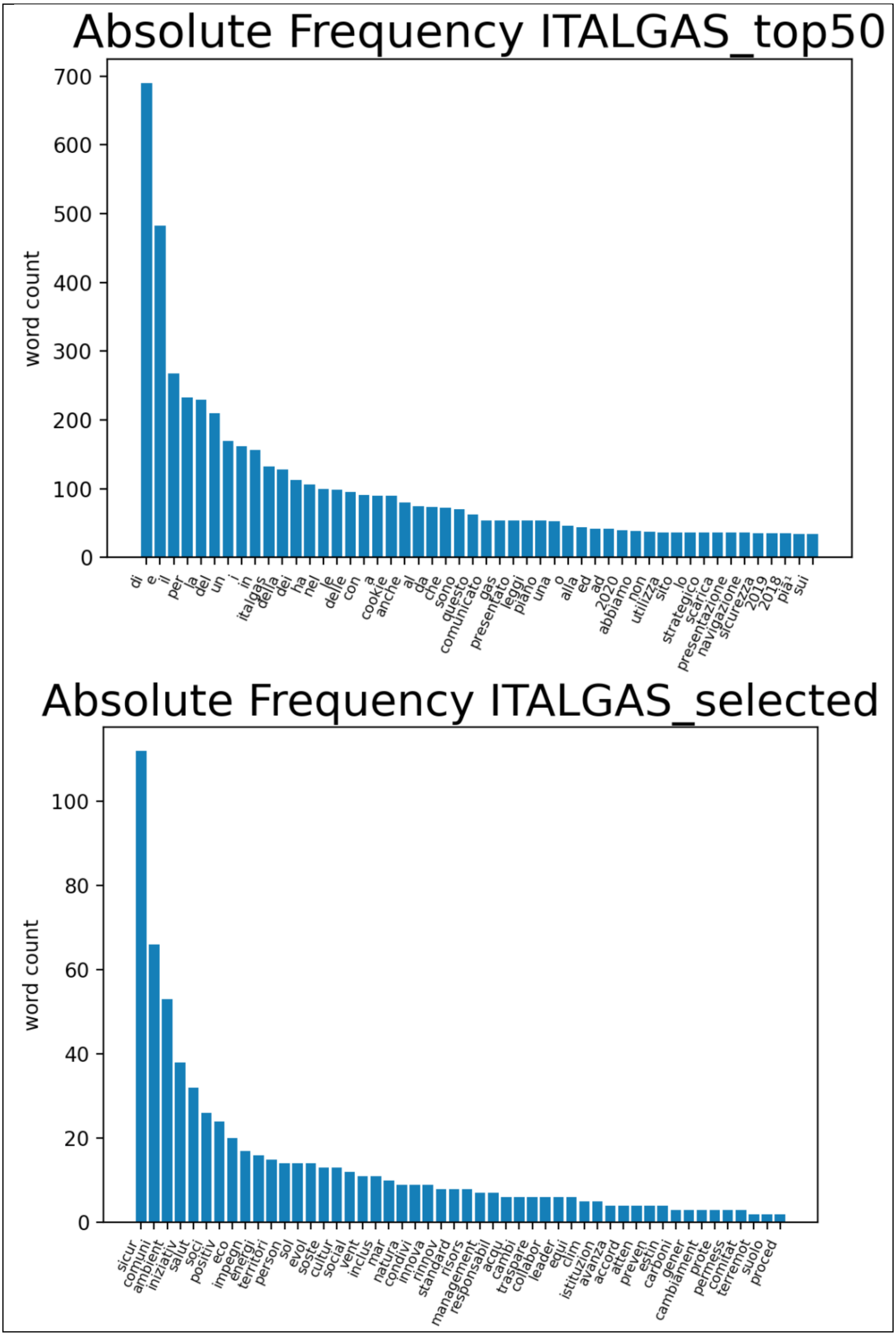
INTESA SANPAOLO



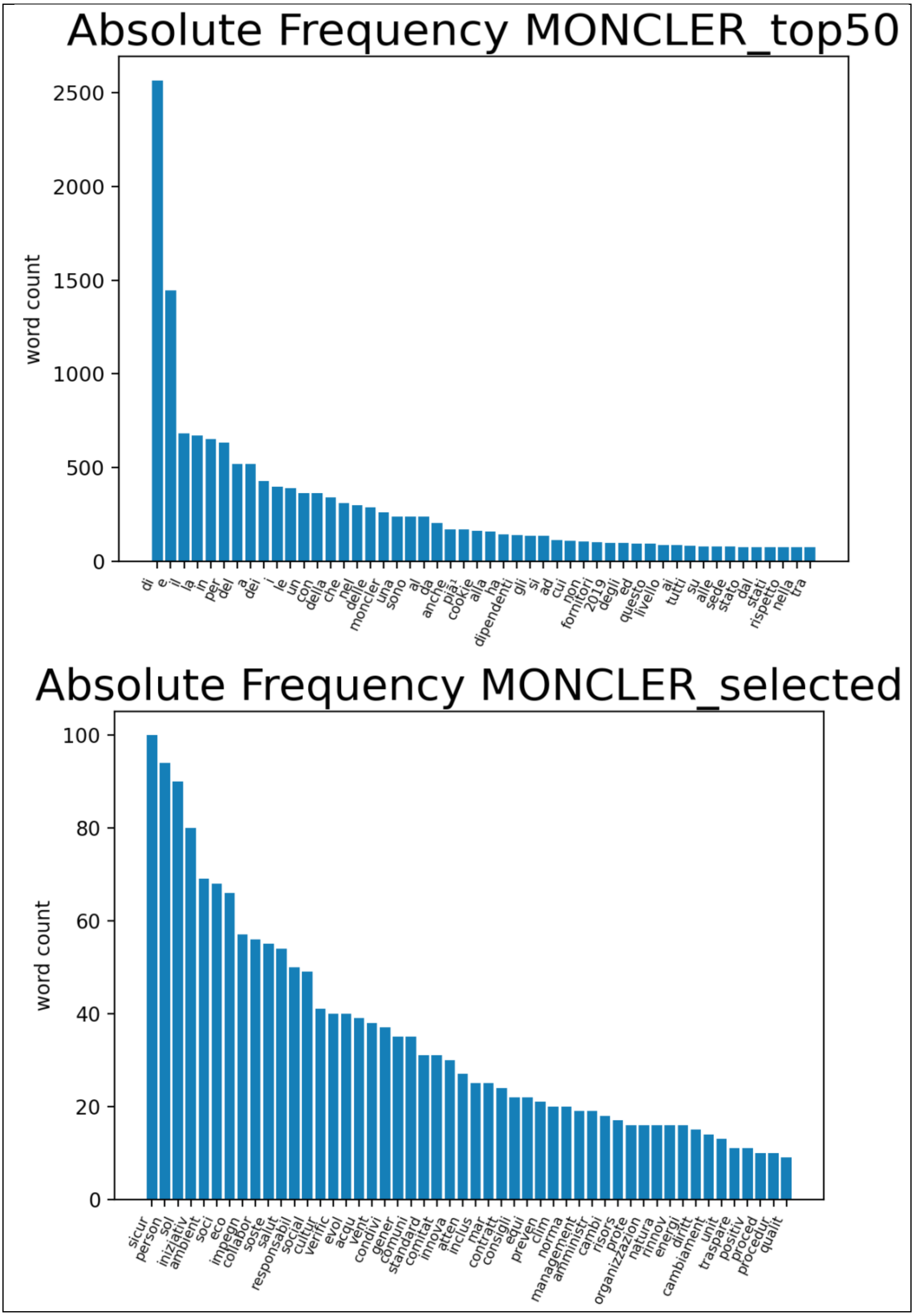
IREN


Absolute Frequency IREN_top50
word count
0
50
100
150
200
250
300
350
di
e
il
del
per
la
in
gruppo
dei
a
con
le
al
iren
nel
delle
della
provincia
un
ha
che
milioni
2019
sono
gli
alla
nella
ultimi
sviluppo
degli
i
euro
piã¹
ai
circa
rispetto
negli
impianti
crescita
riduzione
anni
nei
agli
raccolta
si
cui
grazie
dal
quanto
una
Absolute Frequency IREN_selected
word count
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
territori
soci
acqu
sicur
ambient
social
risors
impegn
amministr
person
innova
soste
consigli
sol
eco
energi
carboni
rinnov
gener
preven
comitat
evol
comuni
istituzion
unit
iniziativ
atten
vent
natura
mar
riscaldament
collabor
cultur
traspare
equi
norma
accord
procedur
proced
cambi
salvaguard
biodegrad
verd
minor
cambiament
prote
salut
condivi
positiv
qualit

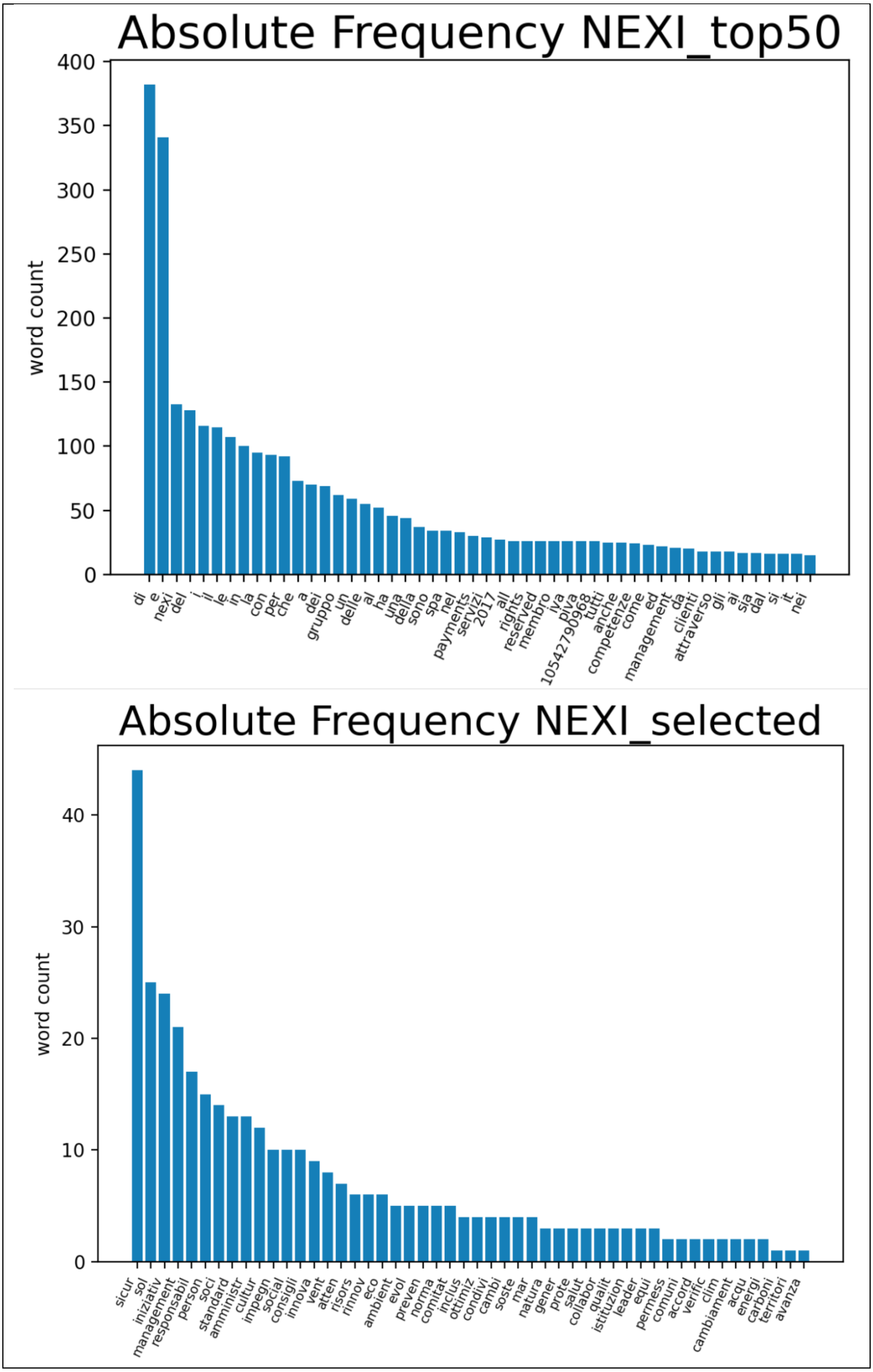
ITALGAS


Absolute Frequency ITALGAS_top50
word count
0
100
200
300
400
500
600
700
di
e
il
per
la
del
un
i
in
italgas
della
dei
ha
nel
le
delle
con
a
cookie
anche
al
da
che
sono
questo
comunicato
gas
presentato
leggi
piano
una
o
alla
ed
ad
2020
abbiamo
non
utilizza
sito
lo
strategico
scarica
presentazione
navigazione
sicurezza
2019
2018
piã¹
sui
Absolute Frequency ITALGAS_selected
word count
0
20
40
60
80
100
sicur
comuni
ambient
iniziativ
salut
soci
positiv
eco
impegn
energi
territori
person
sol
evol
soste
cultur
social
vent
inclus
mar
natura
condivi
innova
rinnov
standard
risors
management
responsabil
acqu
cambi
traspare
collabor
leader
equi
clim
istituzion
avanza
accord
atten
preven
estin
carboni
gener
cambiament
prote
permess
comitat
terremot
suolo
proced

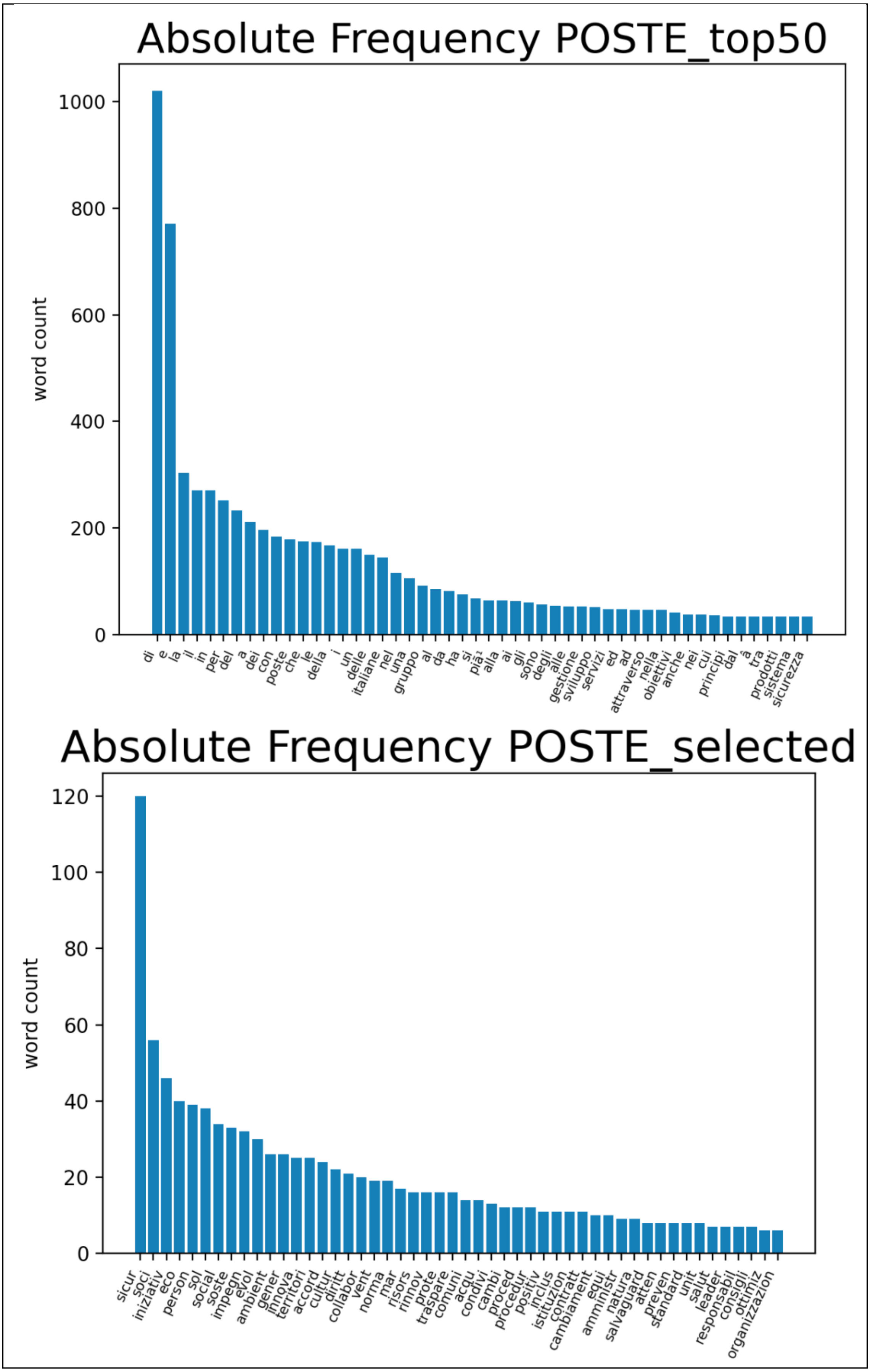
MONCLER


Absolute Frequency MONCLER_top50
word count
2500
2000
1500
1000
500
0
di
e
il
la
in
per
del
a
dei
i
le
un
con
della
che
nel
delle
moncler
una
sono
al
da
anche
piã¹
cookie
alla
ha
dipendenti
gli
si
ad
cui
non
fornitori
2019
degli
ed
questo
livello
ai
tutti
su
alle
sede
stato
dal
stati
rispetto
nella
tra
Absolute Frequency MONCLER_selected
word count
100
80
60
40
20
0
sicur
person
sol
iniziativ
ambient
soci
eco
impegn
collabor
soste
salut
responsabil
social
cultur
verific
evol
acqu
vent
condivi
gener
comuni
standard
comitat
innova
atten
inclus
mar
contratt
consigli
equi
preven
clim
norma
management
amministr
cambi
risors
prote
organizzazion
natura
rinnov
energi
diritt
cambiament
unit
traspare
positiv
proced
procedur
qualit

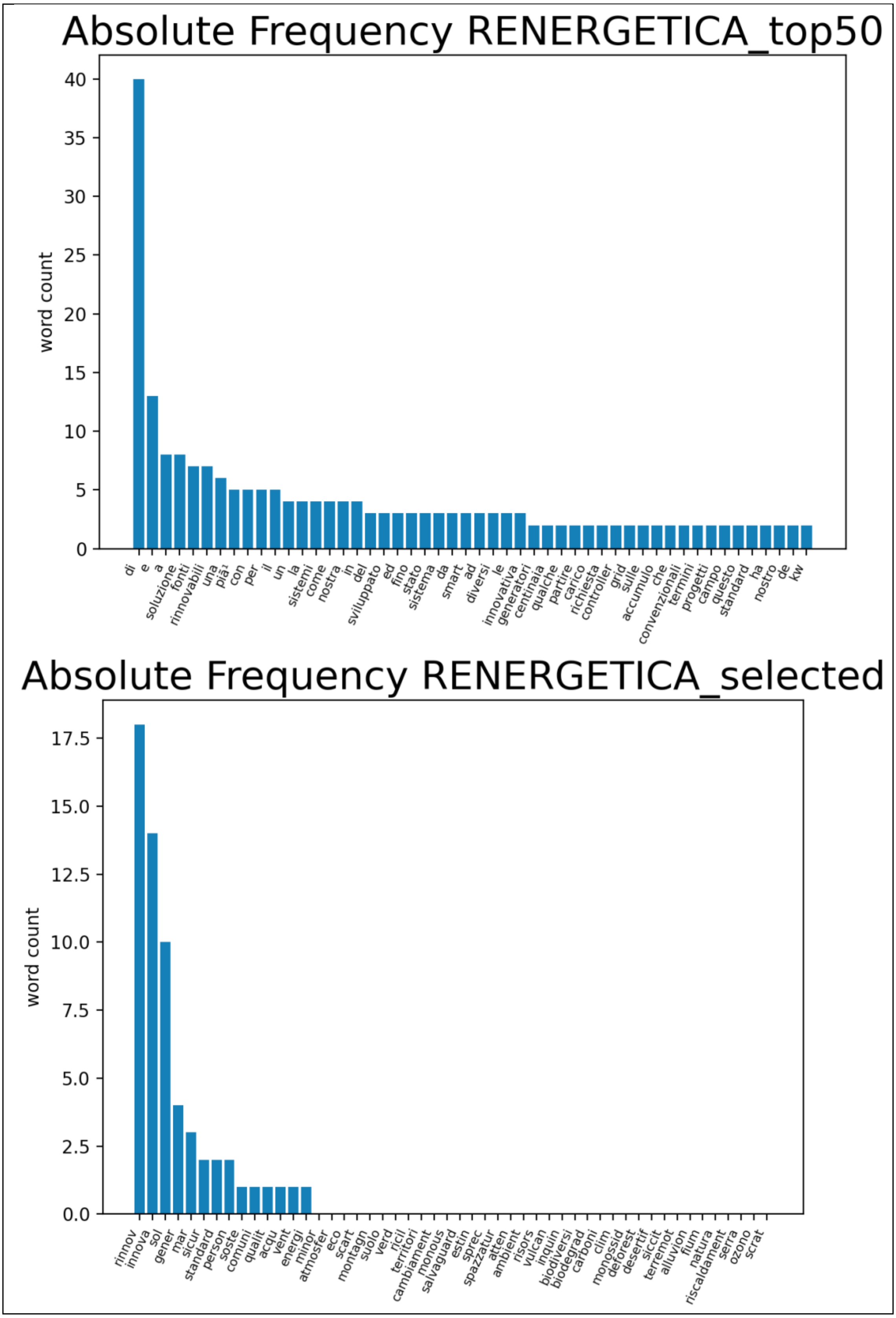
NEXI


Absolute Frequency NEXI_top50
word count
400
350
300
250
200
150
100
50
0
di
e
nexi
del
i
il
le
in
la
con
per
che
a
dei
gruppo
un
delle
al
ha
una
della
sono
spa
nel
payments
servizi
2017
all
rights
reserved
membro
iva
piva
10542790968
tutti
anche
competenze
come
ed
management
da
clienti
attraverso
gli
ai
sia
dal
si
it
nei
Absolute Frequency NEXI_selected
word count
40
30
20
10
0
sicur
sol
iniziativ
management
responsabil
person
soci
standard
amministr
cultur
impegn
social
consigli
innova
vent
atten
risors
rinnov
eco
ambient
evol
preven
norma
comitat
inclus
ottimiz
condivi
cambi
soste
mar
natura
gener
prote
salut
collabor
qualit
istituzion
leader
equi
permess
comuni
accord
verific
clim
cambiament
acqu
energi
carboni
territori
avanza

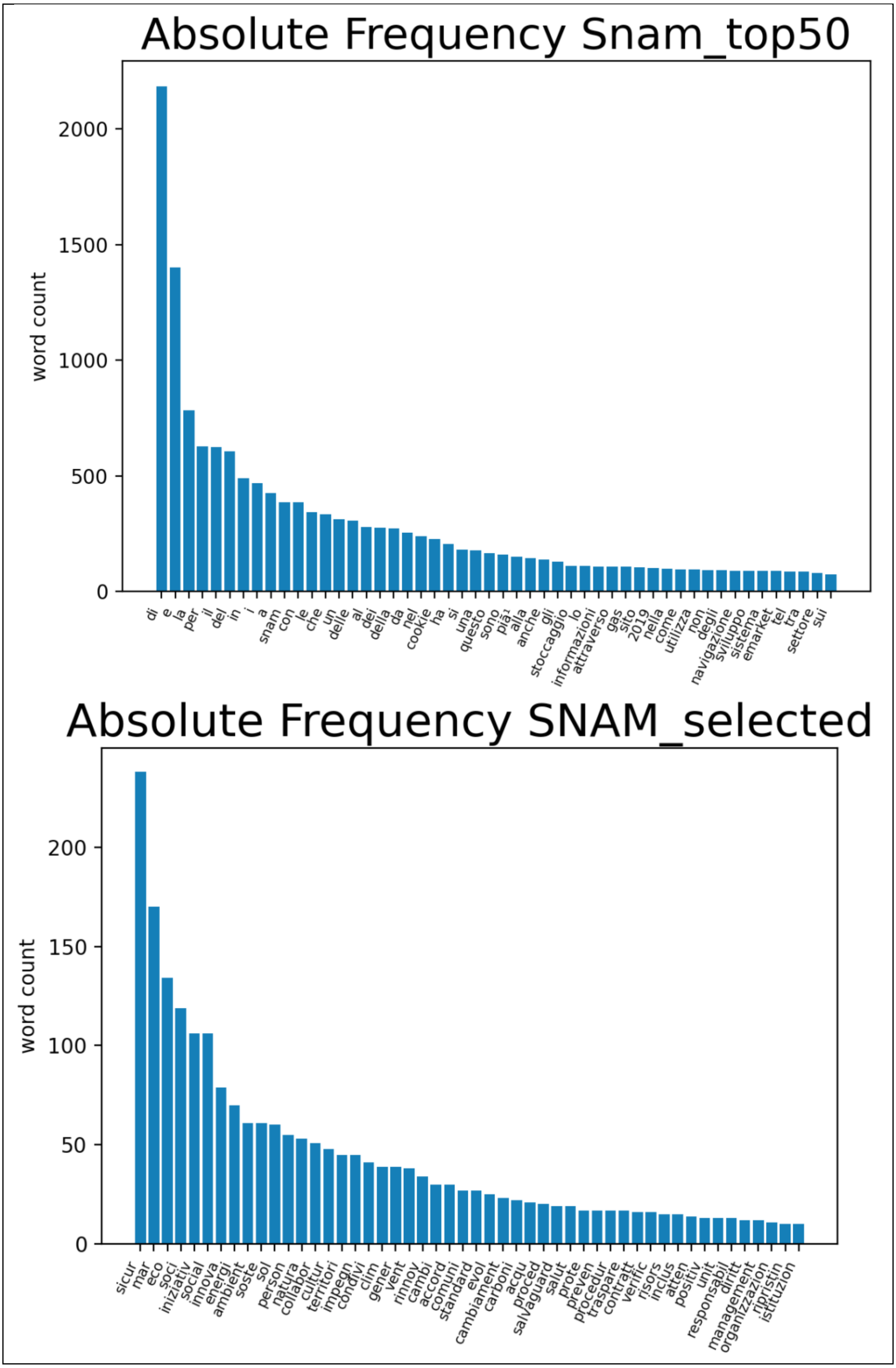
POSTE ITALIANE



RENERGETICA

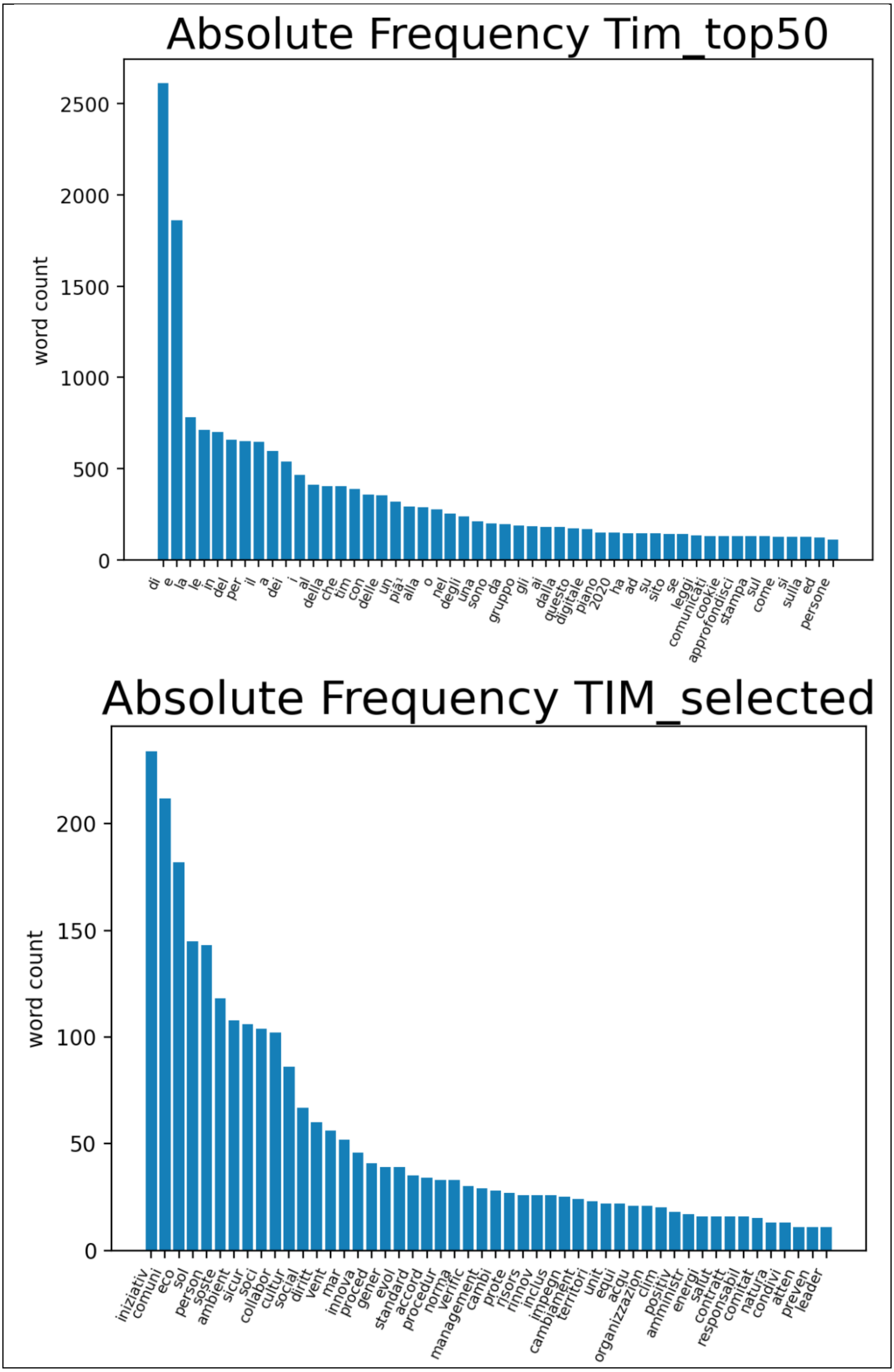


SNAM

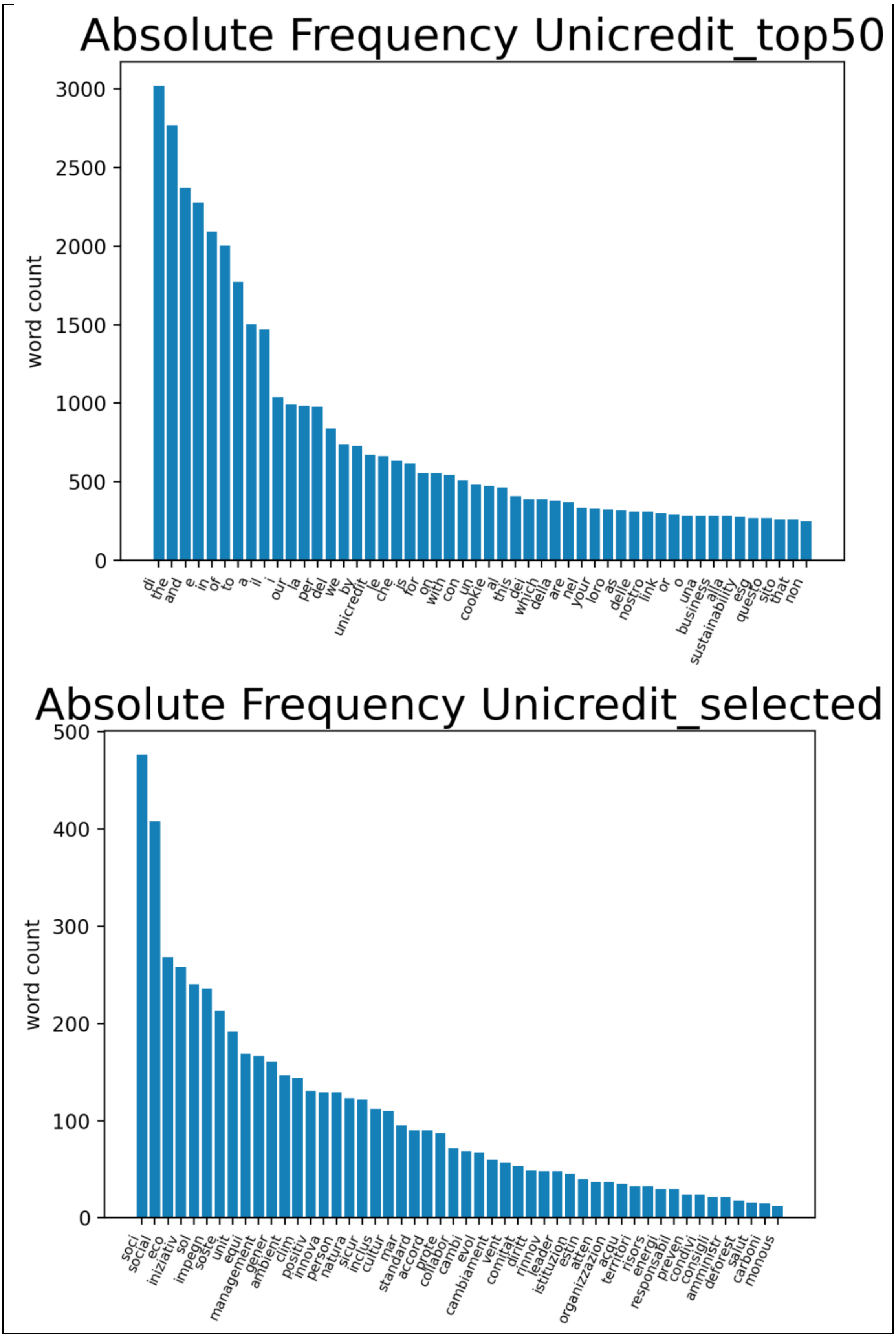


Absolute Frequency Snam_top50
word count
0
500
1000
1500
2000
di
e
la
per
il
del
in
i
a
snam
con
le
che
un
delle
al
dei
della
da
nel
cookie
ha
si
una
questo
sono
piã¹
alla
anche
gli
stoccaggio
lo
informazioni
attraverso
gas
sito
2019
nella
come
utilizza
non
degli
navigazione
sviluppo
sistema
emarket
tel
tra
settore
sui
Absolute Frequency SNAM_selected
word count
0
50
100
150
200
sicur
mar
eco
soci
iniziativ
social
innova
energi
ambient
soste
sol
person
natura
collabor
cultur
territori
impegn
condivi
clim
gener
vent
rinnov
cambi
accord
comuni
standard
evol
cambiament
carboni
acqu
proced
salvaguard
salut
prote
preven
procedur
traspare
contratt
verific
risors
inclus
atten
positiv
unit
responsabil
diritt
management
organizzazion
ripristin
istituzion

TIM


Absolute Frequency Tim_top50
word count
2500
2000
1500
1000
500
0
di
e
la
le
in
del
per
il
a
dei
i
al
della
che
tim
con
delle
un
piã¹
alla
o
nel
degli
una
sono
da
gruppo
gli
ai
dalla
questo
digitale
piano
2020
ha
ad
su
sito
se
leggi
comunicati
cookie
approfondisci
stampa
sul
come
si
sulla
ed
persone
Absolute Frequency TIM_selected
word count
200
150
100
50
0
iniziativ
comuni
eco
sol
person
soste
ambient
sicur
soci
collabor
cultur
social
diritt
vent
mar
innova
proced
gener
evol
standard
accord
procedur
norma
verific
management
cambi
prote
risors
rinnov
inclus
impegn
cambiament
territori
unit
equi
acqu
organizzazion
clim
positiv
amministr
energi
salut
contratt
responsabil
comitat
natura
condivi
atten
preven
leader

UNICREDIT


Absolute Frequency Unicredit_top50

Absolute Frequency Unicredit_selected

Appendix E

Specific topic word evaluation

The code counts the words presented in each subgroup and create a total score for each specific topic: E, S, G.

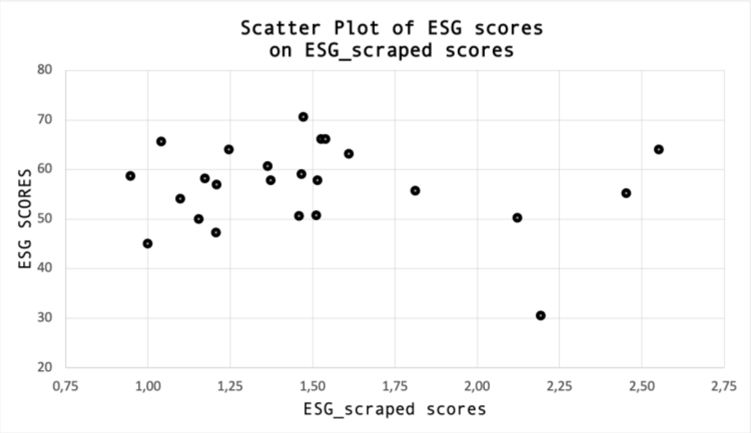
```
import re
import operator
import tkinter as tk
from tkinter import filedialog
import csv

E_words = "biodiversi, biodegrad, carboni, clim, monossid, deforest, desertif, siccit, terremot, energi,
alluvion, mar, fium, risors, natura, riscaldament, serra, rinnov, ozono, inquin, vulcan, spazzatur,
ambient, atmosfer, eco, scart, montagn, vent, sol, acqu, suolo, verd, minor, gener, territori"
S_words = "cambiament, monous, salvaguard, estin, sprec, ricil, atten, scrat, combatt, conserv, prote,
evol, permess, preven, cambi, riform, sicur, ripristin, benefic, salut, avanza, verific, soste, proced,
iniziativ, procedur, ottimiz, rinnov, sicur, standard, traspare, collabor, condivi, impegn, positiv,
innova, qualit, inclus"
G_words = "accord, etic, diritt, dover, socioeconomic, civil, social, comunit, unit, riform, istituzion,
organizzazion, leader, soci, norma, equi, comuni, accord, contratt, comitat, responsabil, maternit,
paternit, management, consigli, amministr, iniziativ, cultur, person"

total_E_words = 0
total_S_words = 0
total_G_words = 0

def get_csv_file():
        root = tk.Tk()
        root.withdraw()

        file_path = filedialog.askopenfilename()
        return file_path

filecsv = get_csv_file()

# Open the file in read mode
with open(filecsv, "r", encoding='latin1') as boh:
  wow = csv.reader(boh)# creating a csv reader object
  for row in wow:
    if row[0] in E_words.split(', '):
          total_E_words = total_E_words + int(row[1])
    elif row[0] in S_words.split(', '):
          total_S_words = total_S_words + int(row[1])
    elif row[0] in G_words.split(', '):
          total_G_words = total_G_words + int(row[1])

print("E =",total_E_words, "S =",total_S_words, "G =",total_G_words)
```

Appendix F

Scatter plots and regression analysis

In this appendix are shown the scatter plots of the variables used in the fourteen regressions and data that are not present in the summary table 3.2.

Regression of ESG scores on ESG scraped scores



SUMMARY OUTPUT

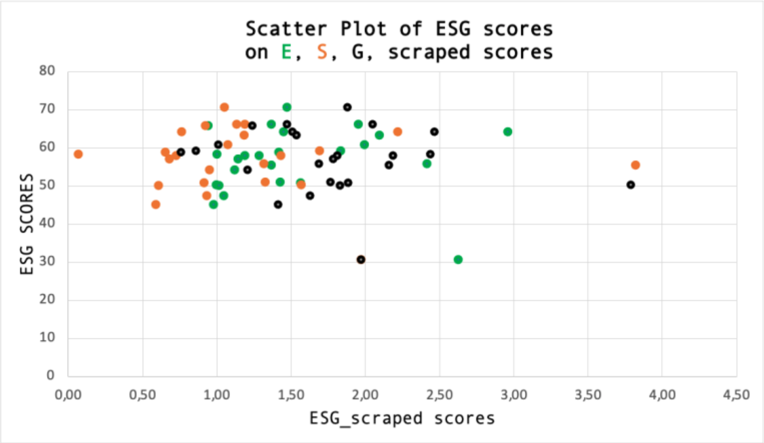
Regression Statistics	
Multiple R	0,13
R square	0,02
Adj. R square	-0,03
Standard Error	8,73
Observations	24

Variance Analysis

	df	SS	MS	F	Significance F
Regression	1	29,66	29,66	0,39	0,54
Residual	22	1676,70	76,21		
Total	23	1706,36			

	Coefficients	Standard Errors	t Stat	P-value	Lower 95%	Upper 95%
Intercept	60,52	6,48	9,34	0,00	47,08	73,96
ESG_scraped scores	-2,59	4,15	-0,62	0,54	-11,20	6,02

Regression of ESG scores on E, S, G, scraped scores



SUMMARY OUTPUT

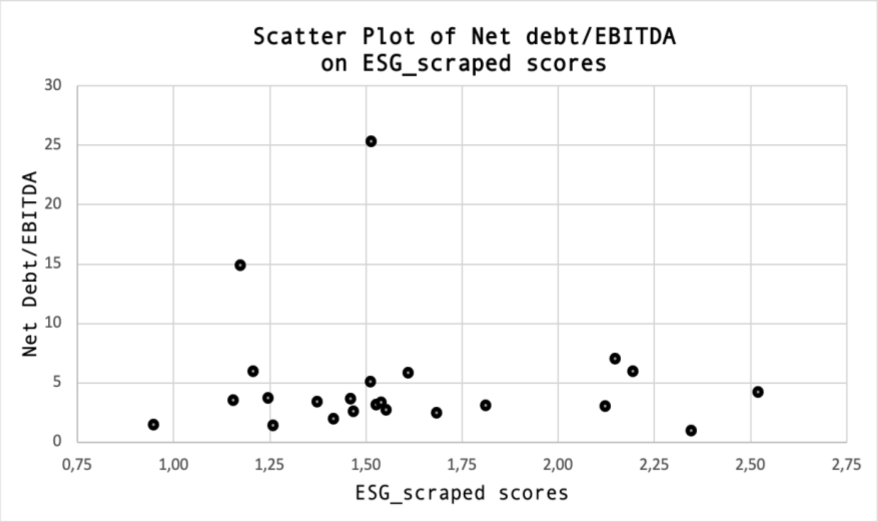
Regression Statistics	
Multiple R	0,18
R square	0,03
Adj. R square	-0,11
Standard Error	9,09
Observations	24

Variance Analisis

	df	SS	MS	F	Significance F
Regression	3,00	55,22	18,41	0,22	0,88
Residual	20,00	1651,14	82,56		
Total	23,00	1706,36			

	Coefficients	Standard Errors	t Stat	P-value	Lower 95%	Upper 95%
Intercept	60,57	8,01	7,56	0,00	43,85	77,29
E_Scra/tot *100	0,49	3,80	0,13	0,90	-7,44	8,42
S_Scra/tot *100	-0,97	2,97	-0,33	0,75	-7,17	5,24
G_Scra/tot * 100	-1,99	3,20	-0,62	0,54	-8,66	4,68

Regression of Net Debt/EBITDA on ESG_scraped scores



SUMMARY OUTPUT

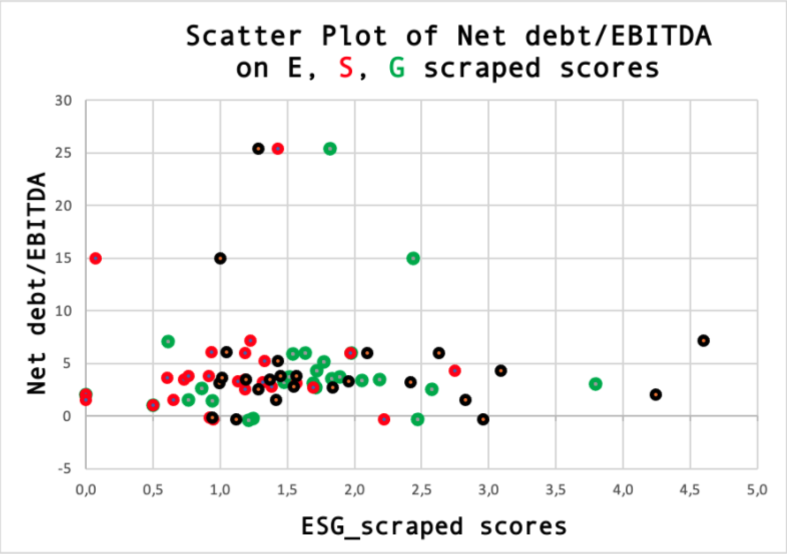
Regression Statistics	
Multiple R	0,05
R square	0,00
Adj. R square	-0,04
Standard Error	5,33
Observations	26

Variance Analysis

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1,92	1,92	0,07	0,80
Residual	24	682,28	28,43		
Total	25	684,20			

	Coefficients	*Standard Errors*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	5,37	3,87	1,39	0,18	-2,62	13,36
ESG_scraped score	-0,61	2,34	-0,26	0,80	-5,43	4,22

Regression of Net Debt/EBITDA on E, S, G, scraped scores



SUMMARY OUTPUT

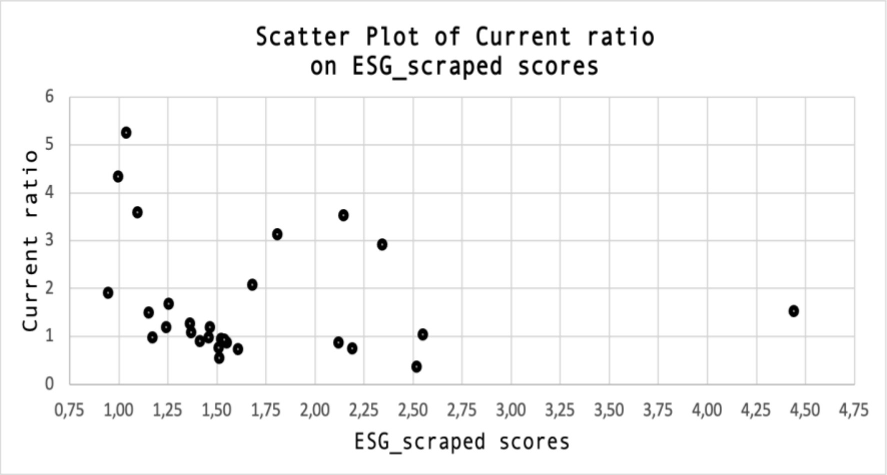
Regression Statistics	
Multiple R	0,20
R square	0,04
Adj. R square	-0,09
Standard Error	5,46
Observations	26

Variance Analysis

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	27,18	9,06	0,30	0,82
Residual	22	657,02	29,86		
Total	25	684,20			

	Coefficients	*Standard Errors*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	4,07	4,48	0,91	0,37	-5,22	13,35
E score	-0,40	1,07	-0,38	0,71	-2,62	1,82
S score	-0,28	1,90	-0,15	0,88	-4,22	3,66
G score	0,91	1,91	0,48	0,64	-3,05	4,86

Regression of Current ratio on ESG_scraped scores



SUMMARY OUTPUT

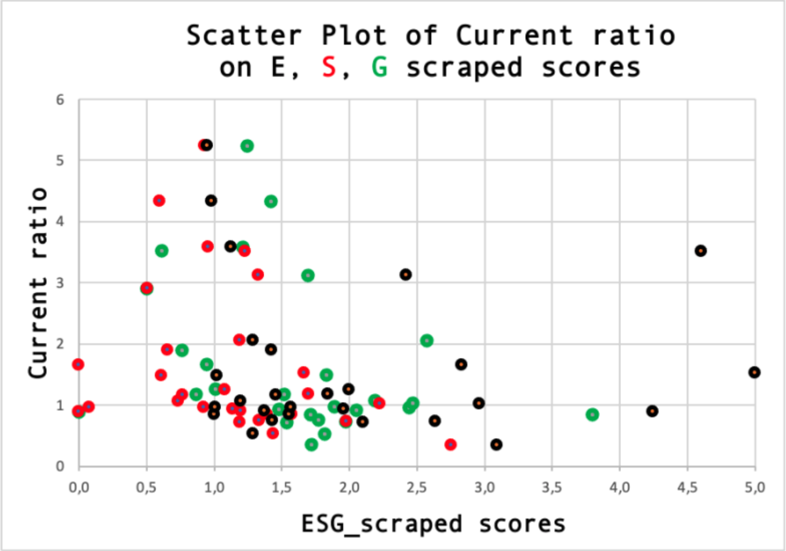
Regression Statistics	
Multiple R	0,18
R square	0,03
Adj. R square	-0,01
Standard Error	1,25
Observations	28

Variance Analysis

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1,34	1,34	0,85	0,36
Residual	26	40,63	1,56		
Total	27	41,96			

	Coefficients	*Standard Errors*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2,19	0,62	3,53	0,00	0,91	3,47
ESG_scraped score	-0,32	0,34	-0,92	0,36	-1,02	0,39

Regression of Current ratio on E, S, G, scraped scores



SUMMARY OUTPUT

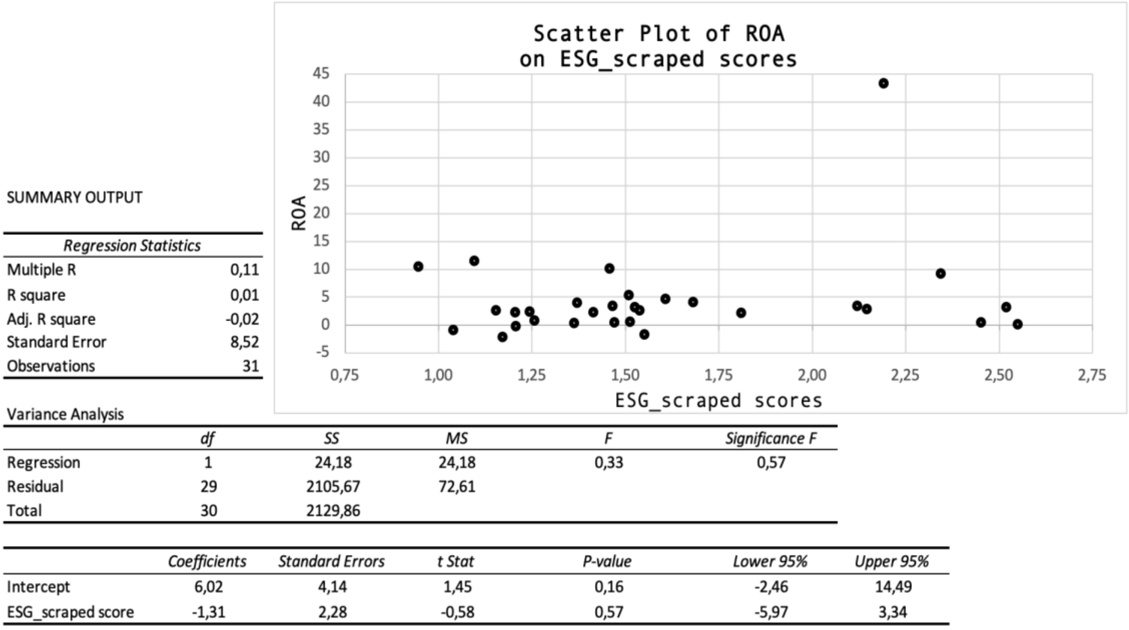
Regression Statistics	
Multiple R	0,31
R square	0,09
Adj. R square	-0,02
Standard Error	1,26
Observations	28

Variance Analysis

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	3,93	1,31	0,83	0,49
Residual	24	38,03	1,58		
Total	27	41,96			

	Coefficients	*Standard Errors*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2,27	0,64	3,55	0,00	0,95	3,60
E score	0,06	0,18	0,33	0,75	-0,31	0,43
S score	-0,44	0,41	-1,08	0,29	-1,29	0,41
G score	-0,14	0,21	-0,66	0,52	-0,58	0,30

Regression of ROA on ESG_scraped scores



SUMMARY OUTPUT

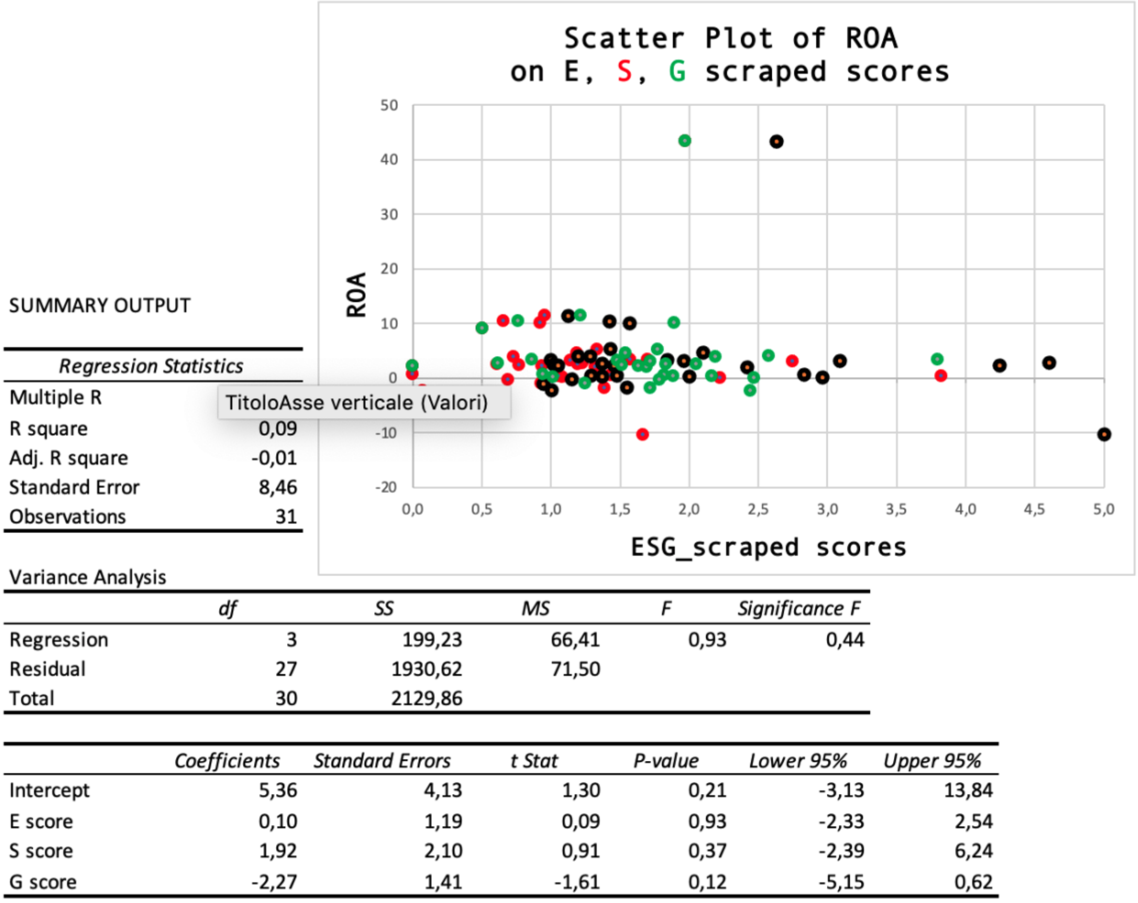
Regression Statistics	
Multiple R	0,11
R square	0,01
Adj. R square	-0,02
Standard Error	8,52
Observations	31

Variance Analysis

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	24,18	24,18	0,33	0,57
Residual	29	2105,67	72,61		
Total	30	2129,86			

	Coefficients	*Standard Errors*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	6,02	4,14	1,45	0,16	-2,46	14,49
ESG_scraped score	-1,31	2,28	-0,58	0,57	-5,97	3,34

Regression of ROA on E, S, G, scraped scores



SUMMARY OUTPUT

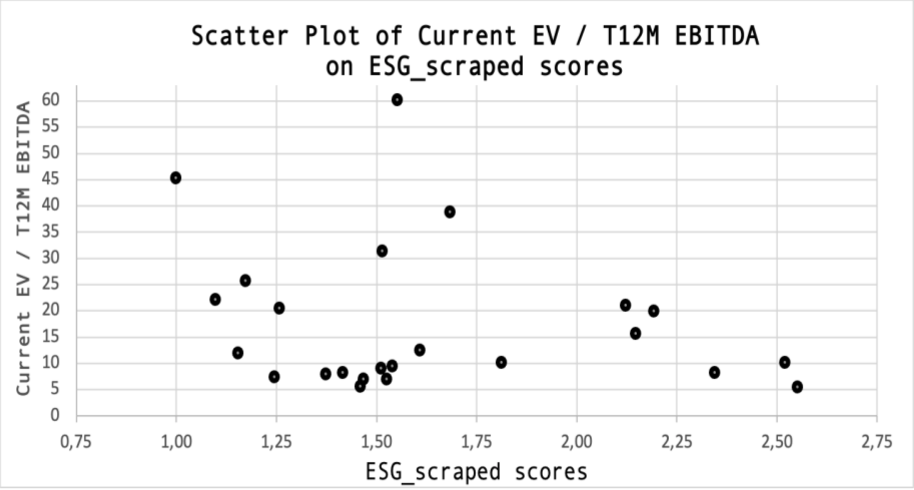
Regression Statistics	
Multiple R	
R square	0,09
Adj. R square	-0,01
Standard Error	8,46
Observations	31

Variance Analysis

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	199,23	66,41	0,93	0,44
Residual	27	1930,62	71,50		
Total	30	2129,86			

	Coefficients	*Standard Errors*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	5,36	4,13	1,30	0,21	-3,13	13,84
E score	0,10	1,19	0,09	0,93	-2,33	2,54
S score	1,92	2,10	0,91	0,37	-2,39	6,24
G score	-2,27	1,41	-1,61	0,12	-5,15	0,62

Regression of Current EV / T12M EBITDA on ESG_scraped scores



SUMMARY OUTPUT

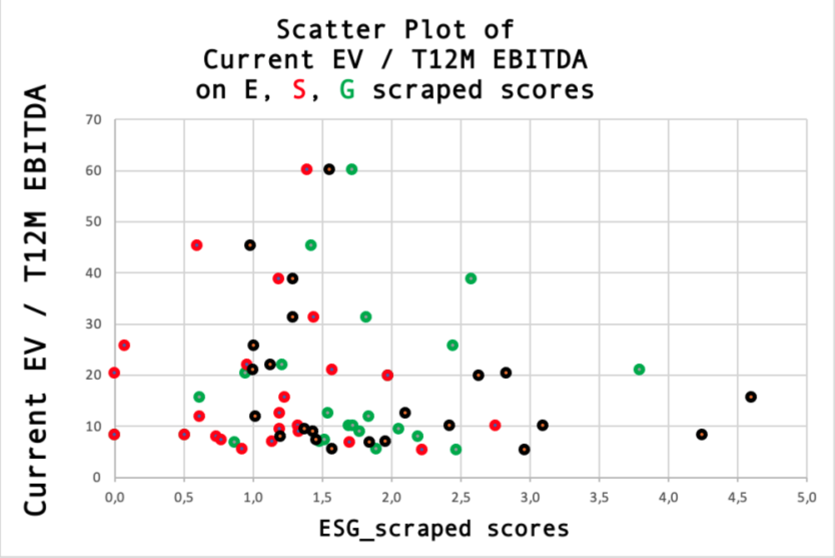
Regression Statistics	
Multiple R	0,24
R square	0,06
Adj. R square	0,01
Standard Error	13,94
Observations	24

Variance Analysis

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	255,98	255,98	1,32	0,26
Residual	22	4274,70	194,30		
Total	23	4530,68			

	Coefficients	*Standard Errors*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	29,67	10,96	2,71	0,01	6,93	52,40
ESG_scraped score	-7,43	6,47	-1,15	0,26	-20,85	5,99

Regression of Current EV / T12M EBITDA on E, S, G, scraped scores



SUMMARY OUTPUT

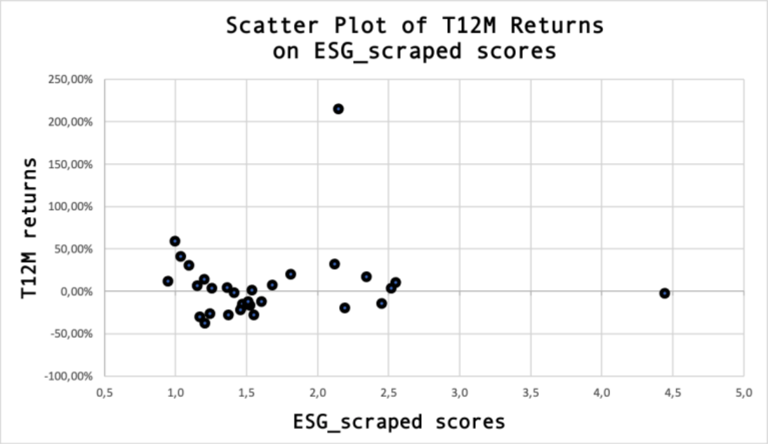
Regression Statistics	
Multiple R	0,33
R square	0,11
Adj. R square	-0,03
Standard Error	14,22
Observations	24

Variance Analysis

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	487,24	162,41	0,80	0,51
Residual	20	4043,44	202,17		
Total	23	4530,68			

	Coefficients	*Standard Errors*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	26,32	13,24	1,99	0,06	-1,29	53,94
E score	-3,42	3,03	-1,13	0,27	-9,75	2,90
S score	-1,56	4,94	-0,32	0,76	-11,86	8,75
G score	0,12	5,42	0,02	0,98	-11,17	11,42

Regression of last 12 month returns on ESG_scraped scores



SUMMARY OUTPUT

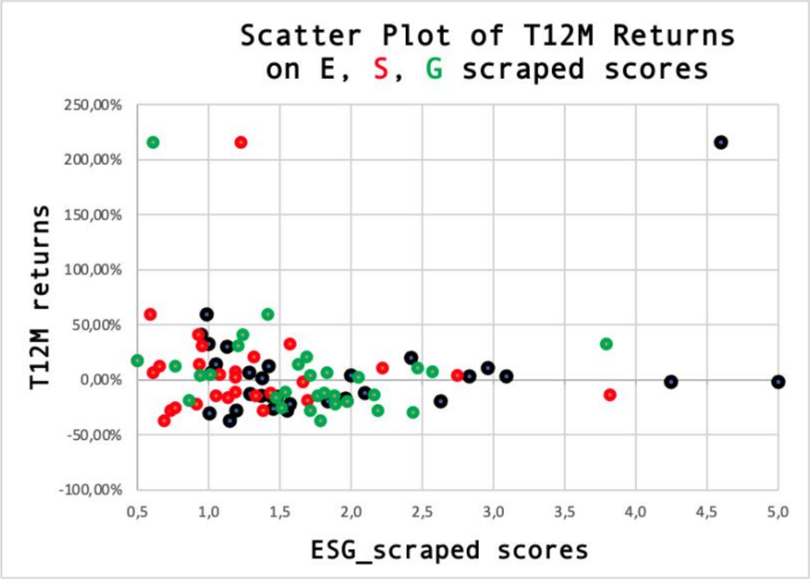
Regression Statistics	
Multiple R	0,09
R square	0,01
Adj. R square	-0,02
Standard Error	0,45
Observations	32

Variance Analysis

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0,05	0,05	0,26	0,61
Residual	30	6,04	0,20		
Total	31	6,09			

	Coefficients	*Standard Errors*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-0,05	0,21	-0,22	0,82	-0,48	0,39
ESG_scraped score	0,06	0,12	0,51	0,61	-0,18	0,30

Regression of last 12 month returns on E, S, G, scraped scores



SUMMARY OUTPUT

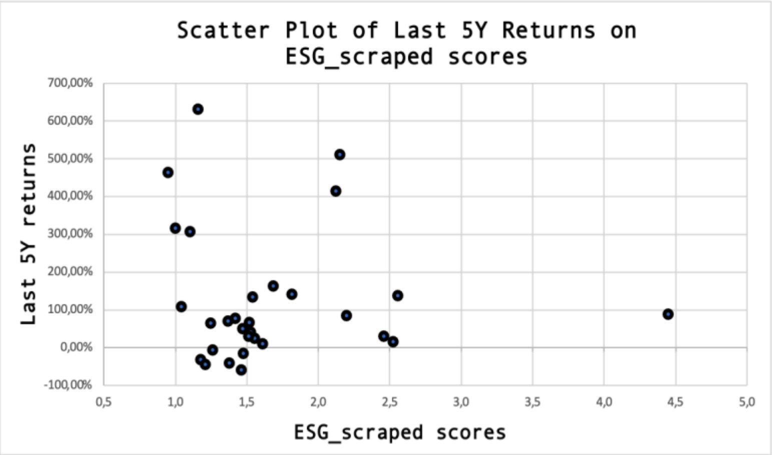
Regression Statistics	
Multiple R	0,38
R square	0,15
Adj. R square	0,06
Standard Error	0,43
Observations	32

Variance Analysis

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	0,90	0,30	1,62	0,21
Residual	28	5,19	0,19		
Total	31	6,09			

	Coefficients	*Standard Errors*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-0,06	0,20	-0,29	0,77	-0,48	0,36
E_score	0,11	0,06	1,91	0,07	-0,01	0,24
S_score	0,02	0,11	0,18	0,85	-0,20	0,24
G_score	-0,08	0,07	-1,12	0,27	-0,23	0,07

Regression of last 5 years returns on ESG_scraped scores



SUMMARY OUTPUT

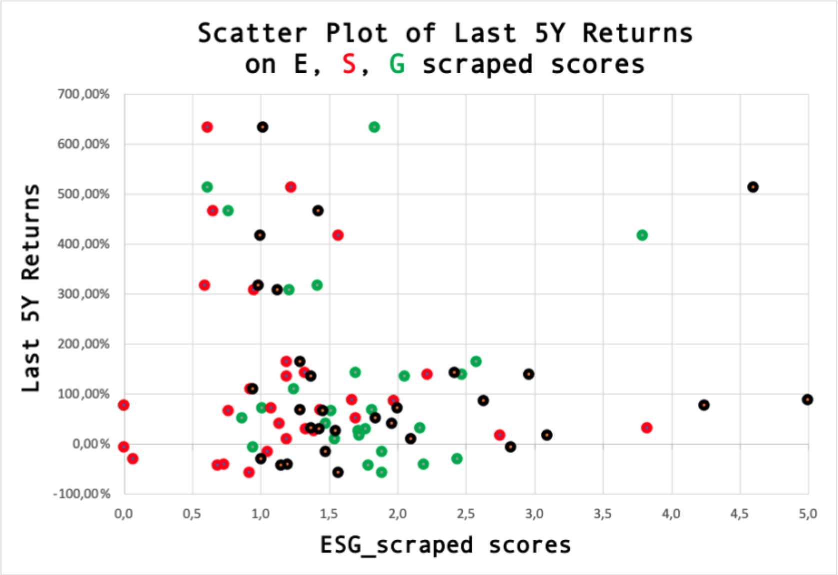
Regression Statistics	
Multiple R	0,067139378
R square	0,004507696
Adj. R square	-0,0310456
Standard Error	1,80014701
Observations	30

Variance Analysis

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0,41	0,41	0,13	0,72
Residual	28	90,73	3,24		
Total	29	91,15			

	Coefficients	*Standard Errors*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1,56	0,87	1,79	0,08	-0,23	3,34
ESG_scraped score	-0,17	0,48	-0,36	0,72	-1,17	0,82

Regression of last 5 years returns on E, S, G, scraped scores



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,10
R square	0,01
Adj. R square	-0,10
Standard Error	1,86
Observations	30

Variance Analysis

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	0,97	0,32	0,09	0,96
Residual	26	90,18	3,47		
Total	29	91,15			

	Coefficients	*Standard Errors*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1,55	0,91	1,71	0,10	-0,32	3,41
E score	0,03	0,32	0,11	0,92	-0,63	0,69
S score	-0,20	0,46	-0,42	0,67	-1,15	0,76
G score	-0,06	0,32	-0,18	0,86	-0,71	0,59

Bibliography

[1] S. A. Mohrman, J. O'Toole and E. E. Lawler, "Attaining corporate social resposnsability practice in corporations: recognising the interdependence of governments and multinational corporations," in *Corporate Stewardship: Achieving Sustainable Effectiveness,* pp. 174-190.

[2] C. Deegan, "Introduction: the legitimising effect of social and environmental disclosures – a theoretical foundation," *Accounting, Auditing and Accountability Journal,* vol. 3, no. 15, p. 282–311, 2002.

[3] C. Deegan, "An overview of legitimacy theory as applied within the social and environmental accounting literature.," *Sustainability Accounting and Accountability,* p. 248–272, 2014.

[4] J. Dowling and J. Pfeffer, "Organisational legitimacy: social values and organisational behaviour," *Pacific Sociological Review,* pp. 122-136, 1975.

[5] USSIF, "Report on US Sustainable and Impact Investing Trends," USSIF, 2020.

[6] E. Avetisyan and K. Hockerts, "The Consolidation of the ESG Rating Industry as an Enactment of Institutional Retrogression," *Business Strategy and the Environment,* vol. 26, no. 3, p. 316–330, 29 June 2016.

[7] F. Berg, R. Rigobon and J. F. Koelbel, "Aggregate Confusion: The Divergence of ESG Ratings," *MIT Sloan School Working Paper 5822-19,* 17 May 2020.

[8] M. Billio, M. Costola, I. Hristova, C. Latino and L. Pelizzon, "Inside the ESG Ratings: (Dis)agreement and Performance," *SAFE Working Paper No. 284, Leibniz Institute for Financial Research,* 15 June 2020.

[9] R. Mitchell, Web Scraping with Python: Collecting More Data from the Modern Web, O' Reilly Media, Inc., 2018.

[10] E. S. Richard and A. M. Richard, Dictionary and introduction to global environmental governance, Earthscan, 2012.

[11] P. Baier, M. Berninger and F. Kiesel, "Environmental, Social and Governance Reporting in Annual Reports: A Textual Analysis," *Financial Markets, Institutions & Instruments, Forthcoming,* 10 September 2020.

[12] A. Ziegler and M. Schröder, "What determines the inclusion in a sustainability stock index? A panel data analysis for european companies," *Ecological Economics,* vol. 69, no. 4, pp. 848-856, 2009.

[13] M. Schwartz, "The “ethics” of ethical investing," *Journal of Business Ethics,* vol. 43, no. 3, pp. 195-231, 2003.

[14] M. Meznar and D. Nigh, "Announcements of withdrawal from South Africa revisited: Making sense of contradictory event study findings," *Academy of Management Journal,* vol. 41, no. 6, pp. 715-730, 1998.

[15] B. Feigenbaum and A. Lowenberg, "South African disinvestment: Causes and effects," *Contemporary Policy Issues,,* vol. 6, no. 4, pp. 105-117, 1988.

[16] G. S. I. Alliance, "Global Sustainable Investment Review 2018," 2019.

[17] B. J. Richardson and W. Cragg, "Being Virtuous and Prosperous: SRI’s Conflicting Goals," *Journal of Business Ethics,* no. 92, p. 21–39, 01 September 2010.

[18] K. Watson, B. Klingenberg, T. Polito and T. G. Geurts, "Impact of environmental management system implementation on financial performance: A comparison of two corporate strategies," *Management of Environmental Quality,* vol. 15, 1 December 2004.

[19] A. Chiarini, "Factors for succeeding in ISO 14001 implementation in Italian construction industry," 10 February 2019.

[20] M.-F. Waxin, S. L. Knuteson and A. Bartholomew, "Outcomes and Key Factors of Success for ISO 14001 Certification: Evidence from an Emerging Arab Gulf Country," *Sustainability 2020,* vol. 12, no. 1, p. 258.

[21] C. Flammer, "Corporate Social Responsibility and Shareholder Reaction: The Environmental Awareness of Investors," *Academy of Management Journal,* vol. 56, no. 3, 19 Jul 2012.

[22] M. Nadeem, T. Suleman and A. Ahmed, "Women on boards, firm risk and the profitability nexus: Does gender diversity moderate the risk and return relationship?," *International Review of Economics & Finance,* vol. 64, pp. 427-442, 29 August 2019.

[23] I. J. Mohammad and N. Rabih, "Board gender diversity and firms' equity risk," *Equality, Diversity and Inclusion,* vol. 36, no. 7, pp. 590-606, 18 September 2017.

[24] B. Scholtens and Y. Zhou, "Stakeholder Relations and Financial Performance," *Sustainable Development,* vol. 16, no. 3, pp. 213-232, 2008.

[25] S. E. Hatane, "Employee Satisfaction and Performance as Intervening Variables of Learning Organization on Financial Performance," *Social and Behavioral Sciences,* vol. 211, p. 619–628, 2015.

[26] J. Eklof, O. Podkorytova and A. Malova, "Linking customer satisfaction with financial performance: an empirical study of Scandinavian banks," *Total Quality Management & Business Excellence,* pp. 1-19, 2018.

[27] J. Nolleta, G. Filis and E. Mitrokostas, "Corporate social responsibility and financial performance: A non-linear and disaggregated approach," *Economic Modelling,* vol. 52, pp. 400-407, January 2016.

[28] V. Lopez, G. Arminda and R. Lazaro, "Sustainable development and corporate performance: A study based on the Dow Jones Sustainability Index," *Journal of Business Ethics,* vol. 75, p. 285–300, 2007.

[29] J. H. Bragdon and J. A. T. Marlin, "Is Pollution Profitable? Environmental Virtue and Reward: Must Stiffer Pollution Controls Hurts Profits?," April 1972.

[30] M. R. Moskowitz, "Choosing Socially Responsible Stocks," *Business & Society Review,* 1972.

[31] S. C. Vance, "Are socially responsible corporations good investment risks?," *Managerial Review,* vol. 64, p. 18–24, 1975.

[32] F. Gunnar, T. Busch and A. Bassen, "ESG and financial performance: aggregated evidence from more than 2000 empirical studies," *Journal of Sustainable Finance & Investment,* vol. 5, no. 4, pp. 210-233, 2015.

[33] G. L. Clark, A. Feiner and M. Viehs, "From the Stockholder to the Stakeholder: How Sustainability Can Drive Financial Outperformance," 5 March 2015.

[34] J. J. Griffin and J. F. Mahon, "The Corporate Social Performance and Corporate Financial Performance Debate: Twenty-Five Years of Incomparable Research," *Business & Society,* vol. 36, no. 1, pp. 5-31, 1997.

[35] J. Endrikat, G. Edeltraud and H. Holger, "Making Sense of Conflicting Empirical Findings: A Meta-Analytic Review of the Relationship Between Corporate Environmental and Financial Performance," *European Management Journal,* vol. 32, no. 5, p. 735–751, 2014.

[36] E. Albertini, "Does Environmental Management Improve Financial Performance? A MetaAnalytical Review-," *Organization & Environment,* vol. 26, no. 4, p. 431–457, 2013.

[37] A. Alshehhi, H. Nobanee and K. Nilesh, "first_pagesettings Open AccessReview The Impact of Sustainability Practices on Corporate Financial Performance: Literature Trends and Future Research Potential," *Sustainability 2018,* vol. 10, no. 2, 2018.

[38] S. Marinko and T. Golja, "Corporate Social Responsibility and Corporate Financial Performance—Is There A Link?," *Ekonomska Istrazivanja-Economic Research,* vol. 25, p. 215–242, 2012.

[39] E. Ortas, J. M. Moneva, R. Burritt and J. Tingey-Holyoak, "Does Sustainability Investment Provide Adaptive Resilience to Ethical Investors? Evidence from Spain," *Journal of Business Ethics,* vol. 124, p. 297–309, 2014.

[40] C. Zhao, Y. Guo, J. Yuan, M. Wu, D. Li, Y. Zhou and J. Kang, "settings Open AccessArticle ESG and Corporate Financial Performance: Empirical Evidence from China’s Listed Power Generation Companies," *Sustainability 2018,* vol. 10, no. 8, 2018.

[41] K. K. Dalal and N. Thaker, "ESG and Corporate Financial Performance: A Panel Study of Indian Companies," *IUP Journal of Corporate Governance* , vol. 18, no. 1, pp. 44-59, 2019.

[42] J.-M. Sahut and H. Pasquini-Descomps, "ESG Impact on Market Performance of Firms: International Evidence," *Management international,* vol. 19, no. 2, p. 40–63, 7 May 2015.

[43] J. R. Evans and D. Peiris, "The Relationship between Environmental Social Governance Factors and Stock Returns," 29 August 2010.

[44] T. M. Fischer and A. A. Sawczyn, "The realtionship between corporate soscial performance and the tole of innovation: evidence form German listed firms. J Manag Contorl, 24, 27–52," *Journal of Management Control,* vol. 24, pp. 27-52, 2013.

[45] P. Velte, "Does ESG performance have an impact on financial performance? Evidence from Germany," *Journal of Global Responsibility,* vol. 8, no. 2, p. 169–178, 2017.

[46] G. Landi and M. Sciarelli, " Towards a more ethical market: the impact of ESG rating on corporate financial performance," *Social Responsibility Journal,* vol. 15, no. 1, pp. 11-27, 2019.

[47] H. Hong and M. Kacperczyk, "The price of sin: The effects of social norms on markets," *Journal of Financial Economics,* vol. 93, no. 1, pp. 15-36 , 2009.

[48] M. Sharfman and C. Fernando, "Environmental Risk Management and the Cost of Capital," *Strategic Management Journal,* vol. 29, pp. 569-592, 2008.

[49] R. Bauer and D. Hann, "Corporate Environmental Management and Credit Risk," *ECCE Working Paper. University Maastricht, The European Centre for Corporate Engagement,* 2010.

[50] S. Chava, "Environmental Externalities and Cost of Capital," *Management Science,* vol. 60, no. 9, 2014.

[51] A. Goss and G. S. Roberts, "The impact of corporate social responsibility on the cost of bank loans," *Journal of Banking & Finance,* vol. 35, no. 7, pp. 1794-1810, 2011.

[52] K. Ye and R. Zhang, "Do Lenders Value Corporate Social Responsibility? Evidence from China," *Journal of Business Ethics,* vol. 104, no. 2, pp. 197-206, 2011.

[53] D. S. Dhaliwal, O. Z. Li, A. Tsang and Y. G. Yang, "Voluntary Nonfinancial Disclosure and the Cost of Equity Capital: The Initiation of Corporate Social Responsibility Reporting," *The Accounting Review,* vol. 86, no. 1, pp. 59-100, 2011.

[54] D. S. Dhaliwal, S. Radhakrishnan, A. Tsang and Y. G. Yang, "Nonfinancial Disclosure and Analyst Forecast Accuracy: International Evidence on Corporate Social Responsibility Disclosure," *The Accounting Review,* vol. 87, no. 3, pp. 723-759, 2012.

[55] S. El Ghoul, O. Guedhami, C. Kwok and D. Mishra, "Does corporate social responsibility affect the cost of capital?," *Journal of Banking & Finance,* vol. 35, pp. 2388-2406, 2011.

[56] R. Albuquerque, Y. Koskinen and C. Zhang, "Corporate Social Responsibility and Firm Risk: Theory and Empirical Evidence," *Management Science,* vol. 65, no. 10, pp. 4451-4469, 2019.

[57] I. Girerd-Potin, S. Jimenez-Garces and P. Louvet, "Which Dimensions of Social Responsibility Concern Financial Investors?," *Journal of Business Ethics,* vol. 121, pp. 559-576, 2014.

[58] L. Li, Q. Liu, D. Tang and J. Xiong, "Media reporting, carbon information disclosure, and the cost of equity financing: evidence from China," *Environmental Science and Pollution Research*, vol. 24, p. 9447–9459, 2017.

[59] BakerMcKenzie, "ESG Regulatory Reform: Impact on Asset Managers," London, July 2019.

[60] European Cimmission, "Action Plan: Financing Sustainable Growth," Brussels, 2018.

[61] European Parliament and The Council, "REGULATION (EU) 2020/852 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL," *Official Journal of the European Union,* 18 June 2020.

[62] European Commission, "REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the establishment of a framework to facilitate sustainable investment," Brussels, 2018.

[63] European Parliament and The Council, "REGULATION (EU) 2019/2088 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 November 2019 on sustainability-related disclosures in the financial services sector," *Official Journal of the European Union,* 2019.

[64] High-Level Expert Group on Sustainable Finance, "FINANCING A SUSTAINABLE EUROPEAN ECONOMY, Final Report 2018," 2018.

[65] ESMA, "Final Report: ESMA's technical advice to the European Commission on integrating sustainability risks and factors in MiFID II," 30 April 2019.

[66] D. R. Woodcock, A. S. Kotte, J. D. Guynn and J. Day, "Managing Legal Risks from ESG Disclosures," *Harvard Law School Forum on Corporate Governance and Financial Regulation,* 12 August 2019.

[67] R. G. Eccles, M. D. Kastrapeli and S. J. Potter, "How to Integrate ESG into Investment Decision-Making: Results of a Global Survey of Institutional Investors," *Journal of Applied Corporate Finance,* vol. 29, no. 4, pp. 125-133, 2018.

[68] C. M. Clarkin, M. Sawyer and J. L. Levin, "The Rise of Standardized ESG Disclosure Frameworks in the United States," *Harvard Law School Forum on Corporate Governance,* 2020.

[69] Task Force on Climate-related Financial Disclosures, "Final Report: Recommendations of the Task Force on Climate-related Financial Disclosures," June 2017.

[70] W. F. Sharpe, "Capital asset prices: A theory of market equilibrium under conditions of risk. 19(3): 425-442," *The journal of finance,* vol. 19, no. 3, pp. 425-442, 1964.

[71] J. Lintner, "The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets," *The review of economics and statistics,* vol. 47, no. 1, pp. 13-37, 1965.

[72] J. Tobin, "Liquidity Preference as Behavior Towards Risk," *Review of Economic Studies,* vol. 25, no. 2, pp. 65-86, 1958.

[73] W. F. Sharpe, "Mutual fund performance," *The Journal of Business,* vol. 39, no. 1, pp. 119-138, 1966.

[74] E. F. Fama and K. R. French, "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics* , vol. 33, no. 1, pp. 3-56, 1993.

[75] O. D. Zerbib, "A Sustainable Capital Asset Pricing Model (S-CAPM): Evidence from Green Investing and Sin Stock Exclusion," *Proceedings of Paris December 2020 Finance Meeting EUROFIDAI - ESSEC,* 2020.

[76] L. Pastor, R. F. Stambaugh and L. A. Taylor, "Sustainable investing in equilibrium," 2020.

[77] D. J. Meyer and J. Meyer, "Relative Risk Aversion: What Do We Know?," *Journal of Risk and Uncertainty,* vol. 31, no. 3, p. 243–262, 2005.

[78] R. Lawson, Web Scraping with Python, Packt Publishing, 2015.

[79] S. J. Mooney, D. J. Westreich and A. El-Sayed, "Epidemiology in the era of big data," *Epidemiology,* vol. 3, no. 26, p. 390, 2015.

[80] J. Bar-Ilan, "Data collection methods on the web for infometric purposes – A review and analysis," *Scientometrics,* vol. 1, no. 50, pp. 7-32, 2001.

[81] J. Postel and J. Reynolds, File Transfer Protocol, 1985.

[82] S. VandenBroucke and B. Baesens, "From web scraping to web crawling. Practical Web Scraping for Data Science," 2018, pp. 155-172.

[83] D. Hoogeveen, L. Wang, T. Baldwin and K. M. Verspoor, "Web forum retrieval and text analytics: A survey," *Foundations and Trends in Information Retrieval,* vol. 1, no. 12, pp. 1-163, 2018.

[84] L. Fu, Y. Meng, Y. Xia and H. Yu, Web content extraction based on Webpage layout analysis, Second International Conference on Information Technology and Computer Science, 2010.

[85] E. Uzun, H. V. Agun and T. Yerlikaya, "A hybrid approach for extracting informative content from web pages," *Information Processing & Management,* vol. 4, no. 49, pp. 928-944, 2013.

[86] X. Yu and Z. Jin, "Web content information extraction based on DOM tree and statistical information," in *2017 IEEE 17th International Conference on Communication Technology (ICCT),* Chengdu, 2017.

[87] P. Bohannon, N. Dalvi and Y. Filmus, "Automatic Web-Scale Information Extraction," in *ACM SIGMOD International Conference on Management of Data,* New York, 2012.

[88] J. Clark, S. Derose and I. Corp, "XML Path Language (Xpath)," in *<Http://Www.W3.Org/TR/Xpath/>,* 1999.

[89] N. Dalvi, R. Kumar, B. Pang, R. Ramakrishnan, A. Tomkins, P. Bohannon, S. Keerthi and S. Merugu, "A Web of Concepts," in *In Proceedings of the Twenty-Eighth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems.,* New York, 2009.

[90] S. O'Reilly, "Nominative fair use and Internet aggregators: Copyright and trademark challenges posed by bots, web crawlers and screen-scraping technologies.," *Loyola Consumer Law Review,* vol. 19, no. 273, 2006.

[91] D. Fisher, D. W. Mcdonald, A. L. Brooks and E. F. Churchill, "Terms of service, ethics, and bias: Tapping the social web for CSCW research.," in *Computer Supported Cooperative Work (CSCW), Panel discussion*, 2010.

[92] A. Demchenko, Is web scraping legal? A comprehensive overview from DataOx, 2020.

[93] M. Gaber, A comprehensive legal guide to web scraping in the US, McCarthy Garber Law, 2020.

[94] P. Baier, M. Berninger and F. Kiesel, "Environmental, social and governance reporting in annual reports: A textual analysis," 2020.

[95] Englisch-hilfen, "Environment – Vocabulary List and Sentences in English," [Online]. Available: https://www.englisch-hilfen.de/en/words/environment.htm. [Accessed 17 Jan. 2020].

[96] MyEnglishPages, "Vocabulary," [Online]. Available: https://www.myenglishpages.com/english/vocabulary-lesson-environment.php. [Accessed 17 Jan. 2020].

[97] RelatedWords. [Online]. Available: https://relatedwords.org/relatedto/social. [Accessed 19 Jan. 2020].

[98] I. qualification, "Corporate Governance - A Basic Glossary," [Online]. Available: https://www.icsa.org.uk/assets/files/pdfs/Policy/GlossaryV6.pdf. [Accessed 13 Jan. 2020].

[99] A. Bodnaruk, T. Loughran and B. & McDonald, "Using 10-K text to gauge financial constraints.," *Journal of Financial and Quantitative Analysis,* vol. 50, p. 623–646, 2015.

[100] D. F. Larcker and A. A. Zakolyukina, "Detecting deceptive discussions in conference calls," *Journal of Accounting Research,* vol. 50, pp. 495-540, 2012.

[101] I. Antonellis and E. Gallopoulos, "Exploring term-document matrices from matrix models in text mining," 2006.

[102] G. Salton, A. Wong and C. S. Yang, "A vector space model for automatic indexing," *Communications of the ACM.*

[103] T. Loughran and B. McDonald, "Textual Analysis in Accounting and Finance: A Survey," *Journal of accounting research*, vol. 54, no. 4, pp. 1187-1230, 2016.

[104] R. Gibson, P. Krueger and P. Schmidt, "ESG Rating Disagreement and Stock Returns," in *European Corporate Governance Institute – Finance Working Paper No. 651/2020*, 2020.

[105] G. Dorfleitner, G. Halbritter and M. Nguyen, "Measuring the level and risk of corporate responsibility – An empirical comparison of different ESG rating approaches," *Journal of Asset Management,* vol. 16, p. 450–466, 2015.

[106] D. A. Sauer, "The impact of social-responsibility screens on investment performance: Evidence from the Domini 400 social index and Domini Equity Mutual Fund," *Review of Financial Economics,* vol. 6, no. 2, pp. 137-149, 1997.