FEDERAL UNIVERSITY OF PERNAMBUCO
CENTER FOR APPLIED SOCIAL SCIENCES
DEPARTMENT OF ACCOUNTING AND ACTUARIAL SCIENCE
GRADUATE PROGRAM IN ACCOUNTING

ARLINDO MENEZES DA COSTA NETO

MACHINE LEARNING AND READABILITY IN ACCOUNTING: AN ENSEMBLE LEARNING APPROACH

Recife

November 2025

ARLINDO MENEZES DA COSTA NETO

Machine Learning and Readability in Accounting: An Ensemble Learning Approach

Thesis presented for the fulfilment of the requirements for the degree of Doctor of Philosophy in Accounting, at the Graduate Program in Accounting, Department of Accounting and Actuarial Sciences, Center of Applied Social Sciences, Federal University of Pernambuco.

Supervisor: Luiz Carlos Marques dos Anjos

Recife

November 2025

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Costa Neto, Arlindo Menezes da Costa.
Machine Learning and Readability in Accounting: An Ensemble Learning Approach / Arlindo Menezes da Costa Neto. - Recife, 2025.
80f.: il.

Tese (Doutorado) - Universidade Federal de Pernambuco, Centro de Ciências Sociais Aplicadas, Programa de Pós-Graduação em Ciências Contábeis, 2025.
Orientação: Luiz Carlos Marques dos Anjos.
Inclui referências e apêndices.

1. Informativeness; 2. Machine Learning; 3. Accounting information. I. Anjos, Luiz Carlos Marques dos. II. Título.

UFPE-Biblioteca Central

Arlindo Menezes da Costa Neto

Machine Learning and Readability in Accounting: An Ensemble Learning Approach

Thesis presented for the fulfilment of the requirements for the degree of Doctor of Philosophy in Accounting, at the Graduate Program in Accounting, Department of Accounting and Actuarial Sciences, Center of Applied Social Sciences, Federal University of Pernambuco.

Thesis defended in November 26, 2025:

Luiz Carlos Marques dos Anjos
Supervisor

Cássio da Nóbrega Besarria
External Examiner

Paulo Salgado Gomes de Mattos Neto
External Examiner

Daniel José Cardoso da Silva
Internal Examiner

Wilton Bernardino da Silva
Internal Examiner

Recife
November 2025

This dissertation is dedicated to Suerda Maria de Menezes Araújo Lima.

ACKNOWLEDGEMENTS

I thank those who enabled, accompanied, and met me during this period of my life. Although my own volition led me to doubt my capabilities and resilience, those in my life have kept me on this path, and for them I must be grateful.

To my supervisor, Luiz Carlos Marques dos Anjos, for believing in me and my ideas, lighting the dark path of knowledge path using the light of his experience.

To the Federal University of Pernambuco, namely professors Cláudio Wanderley, Daniel Cardoso and Giuseppe Trevisan for helping me climb the steps on the stairs of knowledge.

To my mother, Suerda Maria de Menezes Araújo Lima, whose teachings and memory has made every sacrifice worth it. To my father, Durval de Araújo Lima, my brother, Xerxes Messias Menezes de Araújo Lima and my future wife, Luanne Santiago Dantas. The support and care you have provided has carried me until this moment.

To my collegues, Hellen Bomfim Gomes Dias, Lucas Allan Diniz Schwarz, Paulo Bandeira de Luna Júnior, Thiago Rios Sena and Wesley Paulo dos Santos, thank you for being the peers whose influence led me to be better. To my friends, Allyson Souza, Bruno Miguel, Guilherme Eglé e Pedro Gabriel, for the companionship.

Lastly, to all those whose paths crossed mine leaving me better for it, I thank you.

"Every company is looking at AI and deciding where it will help them..."
(Warren Edward Buffett)

RESUMO

Este estudo emprega o FinBERT-PT-BR, um modelo de linguagem baseado em transformadores treinado em textos financeiros em português do Brasil, para desenvolver um Índice de Informatividade, concebido para quantificar o valor informacional das divulgações financeiras. O conjunto de dados é composto por 26.804 notas explicativas anuais de 1.152 companhias abertas brasileiras, abrangendo um período de 12 anos (2011–2023). Além o índice, são calculadas as medidas tradicionais de legibilidade, *Flesch-Kincaid Reading Ease*, Índice de *Fog*, Índice *SMOG* e Índice de *Loughran-McDonald*, para cada nota. Em seguida, aplicam-se modelos de aprendizado de máquina (*Random Forest* e *Gradient Boosting*) para avaliar qual dessas métricas de legibilidade melhor representa o índice de informatividade derivado das três dimensões fundamentais: Padronização (*Boilerplateness*), Completude e Densidade. As análises de importância das variáveis nos diferentes modelos indicam que o Índice de *Loughran-McDonald* é o que mais se aproxima da variação do índice de informatividade, sugerindo que ele é a *proxy* mais eficaz para mensurar a legibilidade dos textos financeiros em português. Esse resultado com base em evidência empírica implica mudanças sobre a relação teórica entre complexidade textual e ofuscação informacional sob a ótica da teoria da agência. A pesquisa contribui para a literatura ao integrar modelos de linguagem e técnicas de aprendizado de máquina ao estudo da qualidade das divulgações financeiras em português, um contexto linguístico e regulatório ainda pouco explorado, utilizando um banco de dados extenso. Pesquisas futuras podem ampliar essa abordagem ao incorporar modelos multilíngues, avaliações humanas ou *embeddings* híbridos, de modo a aprimorar e validar o conceito de informatividade.

Palavras-chaves: Informatividade. Aprendizado de Máquinas. Informação contábil. LLM.

ABSTRACT

We expand on the value relevance of accounting information by exploring a new metric for valuing the financial text, to do so we employ a language model (FinBERT-PT-BR) trained in Brazilian Portuguese to develop an Informativeness Index, assigning scores to 26.804 quarterly financial statement notes from 1.152 companies in Brazil over the span of 12 years. As a verification of our model's capability to understand textual data, we calculate the usual readability metrics (Flesch-Kincaid reading ease, Fog index, SMOG index, Loughran-McDonald Index) for all the notes and employ machine learning models to evaluate which readability metric best represents an informativeness index built upon the dimensions of Boilerplateness, Completeness and Density, expecting our proposed metric to be poorly related to the readability metrics. The evaluation of which readability metric is closest to measuring the informativeness of financial text is based on the feature importance, which indicates the best proxy for financial text readability of Portuguese text is be the Loughran-McDonald Index. The Loughran-McDonald Index is the only one with any relevance in the regressors, and as is based on file size, we assume our metric as capable of measuring textual information value better than common readability metrics, while pointing to the Loughran-McDonald to be a reasonable proxy to informational value of financial text. This research innovates by presenting a new method to quantify the informational value of financial information, contributing to value-relevance literature as well as literature of machine learning employment in accounting research, additionally we do so within a not-so-explored field (Portuguese financial information) with a reasonably large dataset. Further research may be needed to combine our proposed model with market-related metrics or human experiments to increase the validity of the metric concept.

Key-words: Informativeness. Machine Learning. Accounting information. LLM.

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
ML	Machine learning
NLP	Natural Language Processing
SHAP	Shapley Additive Explanations
MDI	Mean Decrease in Impurity
MDA	Mean Decrease in Accuraty
OCR	Optical Character Recognition
BERT	Bidirectional Encoder Representations from Transformers
ELMo	Embeddings from Language Model
LM	Language Model
LLM	Large Language Model
DT	Decision Trees
RF	Random Forest
GB	Gradient Boosting
InfoIndex	Informativeness Index
AdaBoost	AdaBoostAdaptative Boosting
GPT	Generative Pre-trained Transformer
CVM	*Comissão de Valores Mobiliários*
B3	*Brasil Bolsa Balcão*
CPC	*Comitê de Pronunciamentos Contábeis*
IFRS	International Financial Reporting Standards

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS

1 Introduction

1.1 Introduction

Market participants vary widely in their knowledge of a company's affairs, and even though corporate managers are expected to act in the best interests of investors and other stakeholders, information asymmetry often persists, as debated by agency theory (Eisenhardt, 1989). Financial statements help bridge this gap by providing standardized accounting data that mitigates asymmetric information. Yet managers may still manipulate how they present such data, prompting regulators such as the International Financial Reporting Standards Foundation (IFRS) to impose increasingly detailed disclosure requirements (Bradbury et al., 2018; A. Cheung & Hu, 2019). Consequently, financial-reporting text has become a focal point for both market participants and researchers alike (Clatworthy & Jones, 2001).

Michelon et al. (2020) classifies three roles of financial reporting: Valuation, which aids investors (Beaver, 1968); Stewardship, useful mitigator of the agency problems (Eisenhardt, 1989); and Accountability, which also addresses agency problems. As such, the growing volume of textual data, has led researchers to search and improve textual document analysis methods (Senave et al., 2023). These methods in turn are employed to better examine relevant information to investors and market participants, such as financial statements (Efretuei et al., 2022; E. Cheung & Lau, 2016), call transcripts (K. Li et al., 2021), 10-K reports (or equivalent) (Fiordelisi & Ricci, 2014; Dyer et al., 2017) among other sources (Fiordelisi & Ricci, 2014). On the company side, the pertinence of text analysis has already influenced how companies structure their files and texts as a policy to mitigate the perception of negative sentiment by machines (Cao et al., 2023).

Our study seeks to contribute to the accounting literature by debating the role readability metrics play in evaluating the informational value of accounting information. We do this by arguing for an informativeness index of our design. Our contribution to the literature is threefold. First, while we understand the role readability metrics have as indicators of information obfuscation (Nadeem, 2021; Bushee et al., 2018; Linsley & Lawrence, 2007) under the agency theory framework, we challenge its theoretical employment as a proxy for assessing the informational content of financial text. Second, we propose an informativeness index built on three dimensions of qualitative information disclosure (Boilerplateness, Completeness and Density) also built upon an agency theory framework. Third, we leverage Machine learning (ML) techniques, namely Gradient Boosting and Random Forest, due to their ensemble approach and capacity to handle non linear, complex data (Friedman, 2002; Bochkay et al., 2023). Our employment of methods built on two approaches is designed to provide us a assurance of accurate predictability, while also using models seen in accounting research (Ranta et al., 2023). We also leverage a BERT model trained in Brazilian portuguese, as a word embedder able to quantify text while understanding word context. The employment of the BERT model leads to the index on 25.804 financial reports from 1.163 Brazilian companies whose information were presented to Brazil's financial market regulator, *Comissão de Valores Mobiliários* (CVM) between the years of 2011 and 2023.

While Artificial Inteligence or ML have seen rapid growth in numerous areas of data analysis (Sarker et al., 2021) mainly due to their capability of "self-directed learning" (Sarker, 2021), the application of these methods within the context of financial text readability or the informational value of accounting qualitative information remains sparse. As such, to the best of our knowledge, our research presents it self as a novel employment of said methods to contribute to the quantification of the informational value of accounting information.

The Brazilian setting is particularly well-suited for this analysis. First, since 2010 Brazil has fully adopted IFRS, ensuring comparability with global standards; however, disclosures are written in Portuguese, providing a unique non-English, emerging-market environment for textual accounting research. Second, the CVM database offers a large and rich dataset: 25.805 firm-quarter reports from 2011 to 2023, enabling robust cross-sectional and temporal analysis of disclosure reports. Third, emerging markets such as as Brazil are characterized by greater information asymmetry and agency conflicts (Leuz, 2010), making the distinction between "readability" (form) and "informativeness" (substance) particularly critical for investors and regulators. Finally, the findings have direct policy implications for Brazilian regulators and standard-setters (CVM, CPC, B3), as the Informativeness Index highlights whether firms rely on boilerplate, omit key topics, or provide dense, decision-useful disclosures. At the same time, the study broadens the global accounting literature by demonstrating how advanced Natural Language Processing techniques can be applied in IFRS, non-English contexts, extending insights beyond the heavily studied U.S. and European settings.

This research analyzes the quarterly financial reports of 1.152 companies whose financial reports were required to be made available to Brazil's CVM, from 2011 to 2023, resulting in 25.804 financial statement reports. The financial statements are converted to text, treated, and subsequently read and interpreted by a neural language model, FinBERT-PT-BR (Santos et al., 2023), which is capable of mimicking a human reader due to its ability to interpret context, specifically financial text context. This interpretation leads to the possibility of scoring the reports under three dimensions, Boilerplateness, Completeness and Density. Boilerplateness measures the boilerplate text repeated over the years with no additional informational value; Completeness measures the coverage of relevant topics by the text, such as Brazil's interpretation of IFRS topics, "CPCs"; lastly, Density measures the explanatory richness by leveraging the capability of the model to understand context. After the measurement dimensions scoring, an index is created and compared with the calculated readability metrics for the analyzed companies through a Random Forest and a Gradient Boosting regressors, where a feature relevance, combined with the Shapley Additive Explanations (SHAP) analysis indicates how well the readability metrics are to predict the informativeness index.

Although prior research has commonly assessed the informativeness of financial reports through market-based results, such as event-studies (Agarwal, 2020; Merkley, 2014), our data do not allow this validation method due to the expected attrition rate, as such we cannot employ market-based results to validate our model. Instead, we compare our Informativeness Index with the most common readability measures, as we expect our Language-model-based metric captures deeper informational properties beyond the superficial linguistic complexity typically measured by readability indices. Our findings suggest that, within the readability metrics usually employed in Accounting research, the Loughran-McDonald Index is the closest to representing the informativeness of a given text. Additionally, future studies may employ our proposed index to evaluate information of publicly listed companies, or alternatively, experiments can be designed to measure if human readers can validate the dimensions scores as resulted from the FinBERT-PT-BR model approach.

We believe that our study advances accounting research by introducing a novel method to assess whether corporate disclosures genuinely enhance decision usefulness or merely comply with formal requirements, through the proposal of an Informativeness Index that shifts the focus from form to substance. By decomposing informativeness into the dimensions of Boilerplateness, Completeness, and Density, we offer a more nuanced instrument to evaluate how disclosures may either illuminate or obscure the underlying economic reality. Furthermore, we build on the insights of Jabarian & Imas (2025); Cao et al. (2023), who highlight the increasing interaction between corporate disclosures and machine readers, as companies increasingly tailor texts to algorithmic processing, while market participants rely on models to evaluate and even generate financial language. This dynamic may create a scenario where both preparers and users of financial information attempt to outpace one another in decoding or engineering textual disclosures. This perspective underscores the timeliness of employing language models as simulated readers while drawing attention to the broader implications of text analysis in the financial domain.

2 Research Background

2.1 Research background

2.1.1 Accounting information, agency and obfuscation

An agency may be summarized as an relationship, that is, a bundle of contracts between two parties, where the principal delegates the work and the agent performing the work (Jensen & Meckling, 1976). However, this relationship may face problems if the agent and the principal have conflicting goals, or even if the principal is unable to verify the acts of its agent (Eisenhardt, 1989). As to mitigate these problems, the principal may incur in monitoring costs (agency costs) to limit the activities of the agent (Jensen & Meckling, 1976), and one of the many shapes this monitoring may take is through disclosure of financial (accounting) information (Leuz & Verrecchia, 2000; Courtis, 1995; Morris, 1987; Holthausen & Leftwich, 1983).

As markets participants perceive and subsequently feel the necessity to reduce the information asymmetry (Akerlof, 1978), regulating bodies have developed legislation to increase both the quality and the quantity of disclosure, trying to improve the overall quality of financial reporting (E. Cheung & Lau, 2016). Yet, this disclosure goes beyond quantitative information, with text-based financial information, narrative disclosures (NDs) growing in relevance in recent years (Hassan et al., 2019; M. Jones & Smith, 2014).

We can observe the intent with this kind of disclosure within accounting regulation. Under the conceptual framework for Financial Reporting presented by the IFRS, certain qualitative characteristics of useful financial information are presented as "fundamental", relevance and faithful representation, while others are described as "enhancing", such as comparability, verifiability, timeliness and understandability (Foundation, 2018). We can observe the intent of the regulators to provide reporting entities with characteristics able to influence decision making, and, by doing so, it is expected that the disclosure information becomes guided by such tenets.

Consequently, researchers attention has been broadly brought towards accounting information disclosure in its many aspects under the agency theory framework (Morris, 1987), with recent research focusing on the communication aspects of financial reports (Hassan et al., 2019), with readability being presented as the dimension that measures the ease (or lack thereof) of conveying information by text (M. Jones & Smith, 2014). One of the most common aspects investigated under readability research is the concept of obfuscation, the tactic of using writing methods that deliberately mask messages (Courtis, 2004). In other words, under an agency problem framework the agent may make use of text that deliberately reduces the impact of certain events, statements or perceptions, writing hard to read text as to reduce the agent's capability of making decisions, reducing the usefulness of the reported information (Hassan et al., 2019). This framing has direct researchers towards a common understanding that less-readable reports are the deliberate result of bad news management (M. J. Jones & Shoemaker, 1994; F. Li, 2008).

This creates a question of whether corporate disclosures truly enhance decision usefulness or simply complies with formal requirements. Prior studies have often relied on readability metrics such as the Fog Index, Flesch–Kincaid, SMOG, or the Loughran–McDonald file size proxy to assess the ease (F. Li, 2008; Bonsall IV et al., 2017). While valuable, these measures are fundamentally rooted in linguistic simplicity rather than in the informativeness of the text. Financial reporting, however, is not aimed at entertainment or stylistic clarity but at conveying relevant and faithfully representative information that supports investment and stewardship decisions (Foundation, 2018). This creates a research gap, between what the readability metrics are able to measure, and what the financial reporting is supposed to deliver. We follow the contributions readability has done in literature as a metric of obfuscation (Hassan et al., 2019; Clatworthy & Jones, 2001; Bradbury et al., 2018; A. Cheung & Hu, 2019; Du & Yu, 2021), and under the same theoretical framework provided by agency theory(Jensen & Meckling, 1976), we evaluate how readability metrics relate to the informational value of NDs. For that, we employ the most common readability metrics, and propose a measurement of the informational value of financial text by proposing an Informativeness Index composed of boilerplateness, completeness, and density to capture disclosure substance.

2.1.2 Readability

Public companies are legally obligated to disclose financial information to shareholders via annual reports. However, not all provided information is easily readable or specific (F. Li, 2008; Dyer et al., 2017). Consequently, this issue has garnered attention from regulatory bodies of financial markets as well as market participants (SEC, 2013; Salehi et al., 2020; E. Cheung & Lau, 2016). Researchers have been investigating this topic for a considerable time. While the value of the information disclosed by companies can be assessed through various metrics, the evaluation of "Access" of a given textual information has been framed under the term *readability* (Smith & Smith, 1971). Readability is defined as the effective communication of valuation-relevant information (Loughran & McDonald, 2014) and may be seen in literature as a metric for assessing the quality of financial disclosure (Chen & Tseng, 2021).

Although there is consensus on the concept of readability, its measurement is not unidirectional. Some of the primary metrics include the Flesch-Kincaid reading ease score, the Gunning Fog index, the SMOG index, and the Loughran-McDonald Index. Researchers have used many of these methods, both in isolation and in combination (Loughran & McDonald, 2014; Hoberg & Lewis, 2017; Chen & Tseng, 2021; F. Li, 2008; Smith & Smith, 1971; Salehi et al., 2020), yet no definitive consensus has been reached on which model better represents the readability of the text, or if the readability is able to convey the informational value of a given text. Due to this multitude of options, and the market-related metrics limitation, we employ commonly employed metrics of readability in Accounting research, to better understand how they relate to our proposed Informativeness Index.

2.1.2.1 Readability metrics

Readability metrics offer financial information users, as well as information generators (accountants and auditors), a way to better assess the comprehensibility of financial disclosures (Barnett & Leoffler, 1979). Although accounting research has used readability as a measurement for various theoretical constructs within accounting information (Efretuei et al., 2022), the main readability metrics used in this context primarily measure linguistic attributes such as document length, word length, and sentence length (Courtis, 1998, 2004). In the following subsections, we introduce the most common readability metrics, due to their prominent usage within accounting research literature.

2.1.2.1.1 Flesch-Kincaid reading ease score

Widely regarded as one of the primary metrics for evaluating reading complexity, the Flesch-Kincaid reading ease score can be measured on a scale that ranges from 30 and below for "scientific journals" (very difficult) to 90 and above for "comic books" (very easy) (Flesch, 1948). This allows for either categorical or continuous variables. Additionally, the Flesch-Kincaid model has been employed in Portuguese-language research (Martins et al., 1996; Silva & Fernandes, 2009).

$$Flesch - Kincaid\ reading\ ease\ score = 206.835 - (1.015 \times ASL) - (84.6 \times \frac{n_{sy}}{n_w}) \quad (2.1)$$

In this model, ASL represents the average sentence length (words/sentences), n_{sy} is the number of syllables, and n_w is the number of words. While the score typically ranges from 0 to 100, it may exceed these limits at both the lower and upper ends of the scale.

2.1.2.1.2 Gunning Fox Index

The Gunning Fog Index (or Fog index) is another commonly used readability metric in financial reporting, considered comparable to the Flesch-Kincaid model in terms of acceptability by government institutions, researchers, and market participants (Gunning, 1952; Loughran & McDonald, 2014; F. Li, 2008).

$$FOG\ Index = 0.4 \times (Words\ per\ sentence + Percent\ Complex\ Words) \quad (2.2)$$

Complex words are generally defined as those with three or more syllables (Hemmings et al., 2020; F. Li, 2008). Most readability indexes operate under the assumption that longer words and sentences decrease the ease of readability (Loughran & McDonald, 2016; Efretuei et al., 2022). The Fog index has been applied in readability research in accounting (Lang & Stice-Lawrence, 2015), suggested by regulators as a possible measure for filed reports (Lundholm et al., 2014), and is frequently used in debates concerning the suitability of readability formulas in accounting (Bonsall IV et al., 2017; Loughran & McDonald, 2014). Fog index scores are directly proportional to text difficulty: the higher the score, the more difficult the text.

2.1.2.1.3 SMOG Index

As an attempt to provide a simpler alternative, Mc Laughlin (1969) developed a metric based on word and sentence length, positing that longer sentences indicate more complex structures, which in turn make a text harder to read.

$$SMOG\ Index = 1.043\sqrt{Number\ of\ polysyllables \times \frac{30}{number\ of\ sentences}} + 3.1291 \quad (2.3)$$

The SMOG Index reflects the number of years of education required to understand a given text. Thus, a higher SMOG index indicates lower readability. It has been widely used in readability research (Chen & Tseng, 2021; Loughran & McDonald, 2016), and as Loughran & McDonald (2016) notes, the SMOG Index can be a more accurate and simpler alternative to the Fog Index.

2.1.2.1.4 Loughran-McDonald Index

Challenging the applicability of the Fog Index to financial information, Loughran & McDonald (2014) argues that one of the Fog Index components is miscalculated and the other is difficult to measure. Instead, they propose that the file size of the 10-K document serves as a simpler and more effective readability metric.

$$Loughran - McDonald\ Index = \log\left(File\ size\right) \tag{2.4}$$

For research not directly related to 10-K filings (such as ours), there may be less incentive to use the Loughran-McDonald Index, as noted by Chen & Tseng (2021). However, as we understand the theoretical underpinnings of the Loughran-McDonald Index, we still use it as an additional research-validated metric, especially one designed to be easily employed in a digital manner.

2.1.3 Financial text informational value

Accounting research has long employed value-relevance methods to evaluate how market participants price accounting information in asset-pricing models (Ball & Brown, 2013). Within this tradition, scholars have examined the economic definition of information, describing it as anything that can influence the outcome of an event, while exploring how standard accounting disclosures affect pricing in conventional settings (Holthausen & Watts, 2001; Beaver, 1968). A recent line of research has turned to readability as a proxy for the usefulness of reports. For example, Ahn et al. (2023) found that more readable financial statements improve the quality of firm-specific information. However, following the critique of Telles & Salotti (2024), readability alone may not capture the true understandability of a document, thus, even within the value-relevance framework, there is room for a more nuanced interpretation of what constitutes an informative accounting text.

Based on our understanding of the informational value of information, we ground our approach in the IFRS Foundation's qualitative characteristics of useful financial information (relevance, faithful representation, comparability, verifiability, timeliness, and understandability) (Foundation, 2018). To that end, we translate these abstract traits into three concrete, measurable dimensions for narrative disclosures, believing the average of them would be the representative of the average informational content of a given financial text. Due to our lack of market-related information for our corpus we explore readability as a comparison parameter for our proposed informativeness index, and following Telles & Salotti (2024) critique, we expect our model to present deeper informational content than that of the readability metrics.

2.1.3.1 Narrative disclosure dimensions

The first dimension, Boilerplateness measures the extent to which a note contains generic or "boilerplate" content that is repeated across periods (Lang & Stice-Lawrence, 2015). By comparing the same narrative across two consecutive reporting dates we capture shifts in boilerplateness; large, unchanged passages signal potential obfuscation by the reporting entity (Carlé et al., 2023; Bushee et al., 2018). The Boilerplateness dimension is calculated by calculating a similarity scale, comprised of the cosine similarity (Xia et al., 2015) for each firm i with the current report $d_{i,t}$ and previous quarter report as $d_{i,t-1}$, as well as the Jaccard similarity (Ji et al., 2013) between the two reports. Cosine similarity is a known methodology to measure the similarity between two documents (Gunawan et al., 2018; Lahitani et al., 2016; Xia et al., 2015). It works by understanding each document as a vector, and using the angle between the two vectors to measure the similarity of both documents (Bochkay et al., 2023; Schütze et al., 2008).

The use of cosine similarity in the accounting context has been criticized (Srivastava, 2023), but research has focused on studying its viability and provided the appropriate setting in which it can be used (Guo, 2022). The overall concept behind its usage in this research is to measure how similar each other words in two subsequent quarterly reports are. Yet, the similarity of meaning do not indicate the overlap of usage, and that requires the employment of a different metric, Jaccard Index or Jaccard Similarity. Jaccard similarity provides us with a complementary similarity metric, an indication of term overlap between two sets of text (Travieso et al., 2024; Bag et al., 2019). In our usage, it is used to indicate not how similar things are, but how much of if is new. Jaccard singularity has also been employed in accounting research (Brown et al., 2023; Johnston & Zhang, 2021; Fontes et al., 2005).

A high cosine similar text, but with low Jaccard similarity, suggests semantic reuse, where different words we used to convey a similar message, yet, a high Jaccard similarity with a high cosine similarity indicates a possible boilerplate text. A low similarity score for both metrics implies novel text.

By employing equal weights on both metrics, we expected that the Boilerplateness metric is as capable of capturing the repetition of words as is the semantic redundancy, with equal importance. The higher the value for our B dimension, the more novel a given text is, the lower the value, the more boilerplate it is.

$$B_{i,t} = 100 - scale(\alpha \times \cos(v_{i,t}, v_{i,t-1}) + \beta \times Jaccard_{i,t}) \tag{2.5}$$

The Completeness dimension assesses the breadth of accounting topics covered in a narrative statement. By mapping each sentence to relevant IFRS elements (e.g., revenue recognition, fair-value measurement), we measure how comprehensively the disclosure addresses the characteristics of relevance, faithful representation, comparability, verifiability, and timeliness. We employ three different metrics to gauge how complete a given disclosure is: $Coverage$, $Balance$ and $ChecklistHitRate$.

$$C_{i,t} = scale(\alpha \times Coverage + b \times Balance + c \times ChecklistHitRate) \tag{2.6}$$

Coverage represents the number of topics covered in a given text. For that, it requires a number of referenced topics, done by clustering. Clustering has been used in many fields of research, and works by organizing data and abstracting an underlying structure (Krishna & Murty, 1999). In our case, is done by topic modeling (Ferri et al., 2021). There is a multitude of methods or modeling topics from text such as, Latent Dirichlet allocation (LDA) (Yang, 2024; Ferri et al., 2021; Blei et al., 2003), Dirichlet compound multinomial (Doyle & Elkan, 2009), BERTopic (Grootendorst, 2022) or k-means (Thiprungsri & Vasarhelyi, 2011). Our approach employs the k-means clustering model trained on the FinBERT-PT-BR embedding, chosen due to the simplicity of the k-mean model and its ample usage (Ahmed et al., 2020; Likas et al., 2003). The k-mean clustering broadly works by grouping data and measuring the distance between the groups (Likas et al., 2003), and its value, topic wise, is when the rate of intra-cluster variance reduction stabilizes., that can be observed by the "elbow test" (Syakur et al., 2018; Bholowalia & Kumar, 2014), additionally we also employ the Silhouette (X. Wang & Xu, 2019; Shahapure & Nicholas, 2020), Davies-Bouldin (Vergani & Binaghi, 2018) and Calinski-Harabasz (X. Wang & Xu, 2019) tests to verify, choosing the median value as the number of topics (K). The *Coverage* metric contributes to the completeness dimension by quantifying the topics (clusters) represented in the reports.

$$Coverage_{i,t} = \frac{Number\ of\ topics\ with\ p_{i,t}^{k} > \tau}{K} \tag{2.7}$$

Balance captures how much "attention" is devoted to the topics mentioned. Our proposed metric derives from the concept of Shannon's Entropy (Shannon, 1948), a known method for measuring information entropy (Liang et al., 2006), or how uncertain a given distribution is (Rényi, 1961). We follow Shannon's Entropy due to its theoretical framework and its previous usage within Accounting research (Abad-Segura et al., 2021; Abdel-Khalik, 1974). In our application, we measure the topic probability vector, measuring how evenly the content of a given report is distributed across all topics, measuring "balance" as in how balanced the disclosure is on its themes. Thus, in our case, a high entropy implies better balance, as the text is more disperse in the topics it covers. The *Balance* metric complements the completeness dimension by ensuring that a given report reflects not only the number of topics it approaches, but also how even the topics are presented.

$$Balance_{i,t} = \sum_{k=1}^{K} p_{i,t}^{k} log(p_{i,t}^{k}) \tag{2.8}$$

ChecklistHitRate measures how many of the topics expected to be disclosed are presented in the text. It measures, for each report, how many items appear using a keyword checklist per disclosure regulation (CPCs). The *CheckListHitRate* metric allows the completeness dimension to measure how well a given report is able to follow accounting disclosure regulation explicitly. Additionally, it should be noted that it does not leverage the embedded text, it is a text-based search for keywords.

$$CheckListHitRate_{i,t} = \frac{Items\ covered}{Total\ items} \tag{2.9}$$

Density measures linguistic compactness (Johansson, 2008). Shorter sentences with fewer jargon terms indicate higher understandability, whereas verbose or convoluted text reduces density scores. Density is the only dimension related to readability employment in an accounting research environment. We employ four different metrics to evaluate how dense a financial report is: $Explain$, $CrossRef$, $ChangeNarr$ and $Params$.

$Explain$ measures the explanatory depth of a sentence in the text, measured by the explanatory cues and prepositions, such as "devido a", "em razão de" (Due to), "mudança" (Change), "estimamos" (estimate). The $Explain$ metric contributes to density by providing a quantifiable measurement of how much attention an text devotes towards explaining its decisions, changes or topics. $CrossRef$ quantifies the explicit cross-references between text and financial information, for instance, if a text comments on the Y value of its intangible assets. $CrossRef$ increases the density's dimension capability of evaluating how connected the text is to the financial information presented. $ChangeNarr$ follows similar logic, by quantifying the changes in narration, or, the frequency of temporal comparison such as year-over-year. The contribution presented by $ChangeNarr$ is similar to that presented by $CrossRef$, yet it differs by quantifying references to different periods. Lastly, $Params$ quantifies the parameter richness, or how common the text refers to assumptions, such as discount rates, hypothesis, assumptions and so on. $Params$ contributes to the density dimension by quantifying the informational compactness of text, as the inclusion of information such as discount rates or hypothesis leads to a higher explanatory value without adding too much length, directly impacting the density of the text.

$$D^k_{i,t} = scale(\delta_1 \times Explain + \delta_2 \times CrossRef + \delta_3 \times ChangeNarr + \delta_4 \times Params) \tag{2.10}$$

The average of the three dimensions is the Informativeness Index (InfoIndex). The methods employed to calculate the dimensions are made possible due to the leveraging of Natural Language Processing and machine learning methods, with the assistance of large language models. We explore more on the choices made over the methods and how they work in the following section.

$$InfoIndex_{i,t} = \frac{(B_{i,t} + C_{i,t} + D_{i,t})}{3} \tag{2.11}$$

2.1.4 Natural Language Processing

Natural Language Processing (NLP) is a field in which natural language is analyzed through computational techniques, enabling communication between humans and computers and facilitating human-like language processing by artificial intelligence (Fisher et al., 2016). NLP allows researchers to analyze text and extract the informational value of linguistic content. Although early research in this field used non-computerized methods, the incorporation of computers has significantly expanded its capabilities (Fisher et al., 2016).

The advances in computational capabilities have enabled "natural language understanding", a subfield that integrates various NLP processes. This has led to computational models that aim to bridge the cognitive gap between simple data processing and complex reasoning and decision-making, a key aspect of artificial intelligence (Chowdhary, 2020). We define artificial intelligence (AI) as computing systems capable of emulating human reasoning and decision-making when solving complex problems (Tung et al., 2004).

Automated textual analysis of corporate disclosures is a contemporary and relevant topic in finance and accounting research (K. Li et al., 2021). This research contributes to the literature by employing NLP in one main application: the Language Representation Model (in our case, FinBERT-PT-BR), a model designed to use NLP processes to interpret text and its surrounding context, functioning as an artificial intelligence that mimics human reading. Additionally, we employ NLP-focused packages such as *spaCy* for the extraction of metrics such as those employed in the calculation of the Density dimension. Moreover, we leverage Machine learning algorithms to better explain the attributes of how the readability metrics relate to the informativeness index.

This study contributes to the NLP literature in accounting and finance by employing novel methods such as third-generation text embedding (Bochkay et al., 2023), which provides deeper insights than second-generation models like *word2vec* (K. Li et al., 2021), due to its superior context understanding, as well as it lack of dependence of a dictionary or vocabulary mapping (Bochkay et al., 2023), allowing for a more natural, human-like, approach to textual analysis. Additionally, the lack of usage of such models in accounting research contribute to the novelty factor of our research design.

2.1.4.1 Ensemble Learning and Machine Learning

The foundation of ensemble learning lies in the understanding that multiple machine learning models, when combined, provide higher-quality predictions as the errors of one model are compensated by another, leading in a improvement of result accuracy quality (Sagi & Rokach, 2018). This research uses ensemble learning in its machine learning component, specifically Random Forest and Gradient Boosting. Random Forest and

Gradient Boosting derive from decision trees (DT), a class of machine learning techniques used for classification and regression tasks (D.-n. Wang et al., 2022). Gradient Boosting, however, also derives from Boosting, a technique whose accurate prediction stems from the combination of multiple not-so-accurate predictors (Schapire & Freund, 2013).

As put by Quinlan (1996), decision trees express inductive inference, the process of moving from concrete examples to general models. Decision trees perform inductive inference by assessing the relative importance of variables (Song & Ying, 2015) by recursively partitioning the input features into smaller subsets based on the values of those features. Each node in the tree represents a feature split, and each leaf node represents a class label or predicted value, and this works recursively identifying optimal points to split observations within the tree until all observations are classified or regressed (Breiman et al., 2017). Decision trees stand out due to their deployment flexibility and ease of interpretation (Sarker, 2021). Despite the accessibility and flexibility (Lee et al., 2022), the decision tree model can be further improved by consequent ensemble designs.

Boosting, however, follows a different approach, by using a combination of "weaker learners" (algorithms designed to provide an error probability slightly less than a random guess) to provide a better result (Ferreira & Figueiredo, 2012). This combination provides a method able to improve the accuracy of learning algorithms with reasonable transparency when compared to "black-box" schemes (Mayr et al., 2014). Firstly introduced with the Adaptative Boosting (AdaBoost) algorithm (Freund & Schapire, 1997), Boosting is seen in different models, such as Gradient Boosting, yet, the methodological roots are the same (Mayr et al., 2014). Boosting algorithms work by simplifying the predictors or classifiers and performing multiple iterations before ensembling the results into a more accurate estimate (Schapire, 2003).

Random Forest integrates multiple decision trees by randomly selecting sample features (Hindman, 2015). By aggregating the predictions from many trees, Random Forests can capture complex patterns in the data that individual trees may miss, while also reducing the tendency of tree models to overfit, especially when handling low observation counts (Bochkay et al., 2023). In contrast, Gradient Boosting combines decision trees and boosting methods. With Gradient Boosting algorithms, the pseudo-residuals of predictions enhance accuracy by reducing bias and variance through a learning rate (Friedman, 2002), that is, it trains each subsequent tree iteratively based on the mistakes made by the previous tree. As the boosting algorithm is susceptible to overfit if not properly designed (Friedman, 2001), both methods should lead to an improved analysis, due to the shortcomings and strengths of each algorithm. Both random forests and boosting are proven estimators in accounting and finance literature (Ranta et al., 2023).

Both machine learning models are designed as regressors, as such they are employed to make predictions, yet, the interpretation of random forest models and gradient boosting models, as black-box models, are not easily interpretable (Adler & Painsky, 2022; Scornet, 2023), with feature importance being one of the main methods to understand the weight of variables in the final prediction Adler & Painsky (2022); Louppe et al. (2013). Yet, we employ SHAP (Mosca et al., 2022) as an additional method to understand how the variables are able to explain the prediction. Coined by Lundberg & Lee (2017), SHAP employs the Shapley values from game theory (Roth, 1988), and it allows for the interpretation of the feature importance in machine learning models due to the increase in algorithmic transparency (Štrumbelj & Kononenko, 2014; Datta et al., 2016).

2.1.4.2 Word embedding, Language Models and BERT

The combination of deep learning techniques, neural networks, and NLP has resulted in language models (Schomacker & Tropmann-Frick, 2021). Language models (LMs) are built on probabilistic models, employing probability to predict words in a sentence (Jurafsky & Martin, 2024). Many models employ these methods for different functions, following distinct approaches. Commercially popular models, such as GPT (Generative Pre-trained Transformer) (Radford et al., 2018), or not-so-commercial models such as ELMo (Embeddings from Language Model) (Peters et al., 2018) and BERT (Bidirectional Encoder Representations from Transformers) (Devlin, 2018) have subtle differences on how they operate, leading to distinct models to distinct tasks.

Despite the distinct uses or designs, language models generally require many steps to understand text, the first being word embedding. Unlike traditional text analysis models that treat each word independently (Loughran & McDonald, 2016; K. Li et al., 2021), word embeddings capture semantic relationships between words by placing them in a continuous vector space where semantically similar words are closer to each other (Mikolov, 2013; Mikolov et al., 2013). The second step required by language models is vector processing as they are to be used within the probabilistic framework of a language model. This processing is achieved in BERT's case through the use of an architecture designed to convert word vectors into probabilities that reflect the meaning of the text, its position within a sentence, and more, depending on the task, this architecture is the Transformer (Vaswani, 2017). Transformers work due to its self-attention mechanism, granting the model the capability of weighing the importance of different words in a sentence based on their mutual relationships, making it adept at handling complex linguistic structures.

BERT's architecture and logic introduced a new approach to NLP tasks. Unlike earlier models that processed text in a unidirectional manner, either left-to-right or right-to-left, BERT employs bidirectional context for every word in the input sequence (Devlin, 2018). This bidirectional context is possible due to a masked language modeling objective, where certain words are replaced randomly with a designated token, which allows the model learning to predict these missing words based on their surrounding contexts. Through extensive pre-training on large corpora (text database), BERT captures intricate linguistic features that can be fine-tuned for various downstream tasks.

An example of a pre-trained model is FinBERT-PT-BR, a Portuguese-trained version developed by Santos et al. (2023), trained specifically for the Brazilian financial context, following the BERT model trained on english corpus and financial information, FinBERT (Huang et al., 2023). BERT's bidirectional context understanding method and the pre-trained Portuguese model focused on financial information should allow FinBERT-PT-BR to read and properly embed large quantities of financial statements while being aware of context, acting as a simulated human reader. While not easily feasible in human-centric experiments, the simulated human approach provides a novel method with distinct possibilities, with literature debate on the topic (Edossa et al., 2024; Engel et al., 2024). As it can be deduced, the employment of a model such as FinBERT-PT-BR removes the need for a predetermined sentiment dictionary or any sort "dictionary approach" (Bochkay et al., 2023) due to the way a BERT model is able to understand and embed the words based on their context, making the NLP implementation easier, albeit computationally intensive.

3 Research Design

3.1 Research design

This research examines how a neural language model, acting as a human reader, measures the informational value of financial reports, and how this informational value relates to the commonly employed readability metrics [1]. The study uses financial statement notes from 1.163 Brazilian companies whose financial reports were made available to CVM over a 12 years period (2011 to 2023), resulting in 25.804, resulting in 24.642 quarterly reports after treatment. The attrition rate is due to the generality dimension, that requires two periods, thus not considering the first available report. In addition, the four most common readability metrics were calculated for each report.

Our approach employs a dataset of previously scored text which provides algorithm-defined input and output variables, indicating the usage of a supervised machine learning model (Ranta et al., 2023). Additionally, as we use the continuous metric of each score index, as opposed to the labeled "reading grade" provided by the models, we opt for a regression strategy (Nielsen, 2022). Given the universe of machine learning models suited to our specifications, we chose to employ algorithms with different approaches, but already deployed in accounting research literature (Zou et al., 2015; Tan et al., 2019; D.-n. Wang et al., 2022). This leads to the choice of Random Forest and Gradient Boosting. While we do not add an stacked model (Pavlyshenko, 2018), we measure how related the readability metrics and the informative index are based on the feature importance (Adler & Painsky, 2022; Louppe et al., 2013) of the variable in each model, alongside a SHAP analysis (Kim & Kim, 2022). Our employment of SHAP (Mosca et al., 2022) is aimed at as an additional method to understand how the variables are able to explain the prediction.

The research is divided into three major stages: First, data gathering and wrangling, including the calculation of readability metrics; Second, the word embedding by FinBERT-PT-BR and Informativeness Index calculation; and Third, the application of Random Forest and Gradient Boosting machine learning models to the text data, with a final comparison to the Informativeness Index.

3.1.1 Data gathering and treatment

Data gathering was automated with the assistance of custom scripts, using data from Brazil's financial market regulator, CVM. No filter was applied to the companies, as we gathered all reports made available by the regulator. Each file contains the quarterly reports from each available company. It should be noted that the our focus document

[1] For a more detailed look at the files, GitHub.

is presented to the regulators by both listed and unlisted companies, following Brazil's legislation, thus increasing the attrition rate of the disclosed companies and the lack of usage of market-related metrics. After data gathering conversion step was necessary to enable large-scale automated text analysis (El-Haj et al., 2020). Therefore, the original files were processed using an array of packages, including Optical Character Recognition (OCR) methods, namely *Pillow, pdf2image, PyMuPDF, pdfplumber* and *Python-tesseract*, to generate text files for use in the subsequent stages of the study. Multiples methods for text conversion were tested, as any conversion method may present artifacts on conversion (broken phrasing, lack of character conversion or broken text in general) the final files employed were chosen after comparison between the conversions presented and the original file. The comparison was conducted by the authors, by comparing a limited amount of reports, with the text file generated after conversion by multiple methods, the method employed (pytesseract-based) was observed to be the most consistent in keeping the document structure, such as phrasing. Additionally, the text files were embedded twice, once with the numerical characters and one without. Nonetheless, the final processing employed the embedding without numbers, as it provided better values and reduced the chance for character vectoring.

After wrangling, the data set consisted of 24.642 quarterly reports statements from 1.163 companies ranging from 2011 to 2023.

The text files were submitted to a different script that used the *textstat* package, allowing for the calculation of the three readability scores employed. As the Loughran-McDonald Index is calculated over the file size, not the file information, a script calculated the index for the already converted text file size of each quarterly report. Despite our research being focused on financial statements notes, as the files are not necessarily representative of the original 10K fillings, we use Brazil's quarterly report as a proxy for the original 10-K file size, *Formulário de Informações Trimestrais* (ITR). Note that we employ the file size for the converted (text) file. For the Informativeness score, the FinBERT-PT-BR model generated a embed for each file, as they were used in the subsequent calculation of the dimensions. The output was stored in a tabular text file with the company name, period, year, month, industry and the values for the Informativeness Index and the four readability metrics. The descriptive statistics are presented in table 1, while the correlation matrix is presented in table 2 followed by the average yearly score for each metric in table 3. Additional tables as well as images are presented in Annex A and B.

Table 1 – Descriptive Statistics

	Flesch Reading Ease	**Gunning Fog Index**	**SMOG Index**	**LM Index**	**Info Index**
Count	24641	24641	24641	24641	24641
Mean	74.42	16.49	13.90	11.69	46.95
Std	14.13	4.67	2.16	0.75	7.78
Min	0.00	5.70	6.48	8.63	10.57
25%	65.46	14.30	12.33	11.21	42.03
50%	75.88	15.90	13.49	11.81	47.68
75%	84.10	18.16	15.20	12.25	52.29
Max	100.00	254.39	24.44	14.43	80.26

As previous stated, and presented in 1, our corpus is comprised of 24.641 financial reports, after the adjustment that the Generality dimension required the previous report to be calculated, resulting in the attrition of the first report for all companies, and as can be seen by 7 and 8, this has remove both the Factoring and Stock Exchange (Holding) sectors, as they only had one observation. The descriptive statistics also attest the different scale for the metrics, but show no critical information pertaining to the wrong calculation of each metric.

Table 2 – Pearson Correlation Matrix

	Info Index	**Flesch Reading Ease**	**Gunning Fog Index**	**SMOG Index**	**LM Index**
Info Index	1.000				
Flesch Reading Ease	0.412	1.000			
Gunning Fog Index	-0.338	-0.689	1.000		
Smog Index	-0.443	-0.861	0.587	1.000	
LM Index	0.668	0.552	-0.334	-0.654	1.000

The correlation matrix shown in 2 indicates the proposed *Info Index* maintains consistent associations with traditional readability metrics. Yet, the presence of a strong positive correlation with the *Loughran-McDonald Index* ($r = 0.668$) points towards the capability of the Index (LM Index) to capture more information than the rest of the metrics. This is expected, as the LM index is derived from the file size, not necessarily from the text-metrics, as such it carries more value than other metrics. In general, these results suggest that the proposed *Info Index* aligns conceptually with established readability constructs, while extending their interpretive scope by emphasizing the informational value of narrative financial disclosures rather than their surface-level complexity alone.

Table 3 – Metric Average - Yearly

Year	Info Index	Flesch-Kincaid_Score	Gunning_Fog_index	SMOG_Index	LM_Index
2011	44.417	71.697	16.829	14.212	11.560
2012	45.896	71.482	17.092	14.213	11.580
2013	45.738	72.201	16.793	14.132	11.615
2014	45.647	72.913	16.637	14.018	11.659
2015	45.608	74.119	16.404	13.900	11.641
2016	45.853	75.321	16.225	13.754	11.650
2017	46.306	75.756	16.193	13.783	11.672
2018	46.972	75.173	16.410	13.890	11.713
2019	47.285	75.692	16.685	13.809	11.726
2020	49.113	73.669	16.919	14.081	11.754
2021	48.657	74.758	16.364	13.967	11.750
2022	48.760	76.030	16.183	13.705	11.791
2023	48.693	77.448	15.807	13.432	11.819

As it can be seen in the average yearly score for each metric, as presented in table 3, there is little change between the score from the first to the last year observed, except when looking at the the Informativeness Index, $Info\ Index$ and The Flesch-Kincaid_Score. Yet, a more in-depth look at the yearly values per sector, as shown in tables 14,15,16,17, for both sectors and holdings, there is no clear trend between the data.

4 Results

4.1 Results

This section presents the results for the machine learning regressors, the model validation procedures, and the analysis of feature importance, as the method to identify which readability metric best predicts the informativeness of Portuguese financial statement notes. The analysis is divided into four complementary stages: First, the assessment of the overall model performance; Second, the verification of model stability through cross-validation; Third, the estimation of variable importance through various means; Fourth, we discuss our findings.

Each stage allows a more comprehensive understanding of how traditional readability indicators relate with a semantic-based informativeness score derived from a large language model (FinBERT-PT-BR) interpretation of financial text informational value.

Our machine learning regressors are performed under a 80/20 split. This indicates that 80% of the data (19.714) of financial statement notes were used as a training dataset, with 20% left (4.928) of the financial statement notes being used to test the trained algorithm. This step is required to verify how the machine learning model is expected to handle new, unobserved data (Bengio et al., 2017).

4.1.1 Model performance

The values presented in table 4 provides the resulting metrics for the Random Forest and Gradient Boosting models. As the models employed are regression-based, metrics such as the Mean Squared Error and the Mean Absolute Error can be used to determine a preferred algorithm. Error measures can explain the difference between the predicted and the observed values within a dataset (Pishro-Nik, 2014) and check for outliers (Chicco et al., 2021). The R^2 (R-squared) value, as indicated by literature, is the main metrics for model adherence within machine learning applications (Chicco et al., 2021). The R-squared both Gradient Boosting and the Stacked models are over 0.5, as we understand R-squared as a measurement of goodness of fit (Cameron & Windmeijer, 1997), we can understand that both the Random Forest and Gradient Boosting regressors moderately represents the variation in the target variable, and while the Gradient Boosting model has a better R-squared value, the difference is marginal. Nonetheless, we can look at the error metrics (MSE, MAE and MAPE) to try to better understand how well the models are able to predict the data. Using the Mean Absolute Percentage Error, MAPE, for both models are below the threshold of 10, indicating a highly accurate forecasting (Moreno et al., 2013; Meade, 1983).

Table 4 – Machine Learning Regressors Summary

	Random Forest	**Gradient Boosting**
$\mathbf{R^2}$	0.5164	0.5299
Mean Absolute Error (MAE)	4.1310	4.1062
Mean Square Error (MSE)	28.5735	27.7699
Root Mean Square Error (RMSE)	5.3454	5.2697
Mean Absolute Percentage Error (MAPE)	9.0392	8.8893

4.1.2 Hyperparameter and Cross-validation tuning

As Hyperparameters are parameters able to affect how a machine learning model learns (Bengio, 2000), we explore changes within the Hyperparameters to obtain the best possible specification. We explore a hyperparameter tuning technique, *RandomizedSearchCV*, to our Gradient Boosting model, allowing for the better fine-tuning of its hyperparameters. The best available specifications were then recorded and used in the model specification (learning rate of 0,04, with a max depth of 7, minimal samples per leaf of 8, minimum samples for split of 4 and 211 estimators). In addition to the hyperparameter optimization for the Gradient Boosting model (chosen due to its better results, while Random Forest was kept as a robustness verification for out main interest, the feature importance), both models underwent k-fold cross-validation. We employ cross-validation as a resampling method for machine learning methods, whose results lead to improved model selection, increasing predictability and reducing overfitting (A. Ramezan et al., 2019). Cross-validation (under the k-fold method, as is our case) works by dividing the sample set into folds (groups), where all but one of the groups is used as test, while the other groups are used as training, this procedure being repeated by the number of folds used (A. Ramezan et al., 2019) (In our case, we explore 5 folds). This leads to an increase in predictive power by the model (Tougui et al., 2021).

Table 5 – Cross-Validation Summary

Model	**Mean $\mathbf{R^2}$**	**Std. dev. ($\mathbf{R^2}$)**	**Range (min-max)**
Random Forest	0.5035	0.0067	0.493-0.514
Gradient Boosting	0.5190	0.0088	0.508-0.533

The values presented in table 5 are the resulting metrics for the Random Forest and Gradient Boosting models after the cross-validation procedure. The results show the effects in the R^2 value. The mean R^2 value after the cross-validation are similar to those before, around the 0.50 range, indicating the previous results were not overfitting, additionally, the Gradient Boosting model keeps a slight advantage over the Random Forest model. The slow standard deviation for both models implies the models are stable even on the subset of the data, and while the advantage is minimal, the Random Forest model appears to be more consistent. In a general manner, it can be understood that the readability metrics are capable of explaining roughly half the variation in the informativeness index, doing so in a consistent manner across groups of data, and while the R^2 indicates a moderate explainability, the low volatility when cross-validation implies the stability of the relationship. Following this validation, we continue to investigate which of the readability metrics better represents Portuguese financial statement text informativeness value.

4.1.3 Feature importance analysis

As stated previously, this research has aimed to verify an alternative method to evaluate which readability metric better represents Portuguese financial statement text. To this end, we employ multiple machine-learning methods to have different approaches to measure the relation between the readability metrics and a developed Informativeness score generated with the help of a Large Language Model, FinBERT-PT-BR.

We employ feature importance as the method for determining which variable (in our case, readability metric) has more impact in the informativeness score predictability. Feature relevance (variable importance) refers to the contribution each input provides to the machine learning algorithms prediction (Hall, 1999, 2000). In our case, the feature importance is directly linked to the research question, directly addressing how a certain readability metric can best predict the informational value of a Portuguese financial text.

There is much debate on how to evaluate feature importance in tree-based models (Scornet, 2023). Following literature we base our findings on the three main methods literature employs, Mean Decrease in Impurity (MDI) or "Impurity-based feature importance) (Scornet, 2023; Disha & Waheed, 2022), Mean Decrease in Accuraty (MDA) or "Permutation-based feature importance"(Altmann et al., 2010) and SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017). The impurity-based feature importance works by going through each node in the random forest and measuring how "pure" (aligned) they are in their prediction, the more a variable (feature) is to making predictions more accurate, the more relevant it is. Permutation-based feature importance however, works by measuring the performance of a mode if one of the features is randomly shuffled. Additionally, we approach feature importance in a different manner, as to increase robustness, with a newer and more theoretically grounded method (Mosca et al., 2022), SHAP. SHAP works by measuring how much a variable is able to sway the prediction away from the average. In our use-case we employ the TreeSHAP model, the better application for tree and ensemble based models (Mosca et al., 2022; Lundberg et al., 2020). The results for the MDI and MDA methods for each model are presented in table 6.

Table 6 – Impurity and Permutation feature importance

Feature	Random Forest		Gradient Boosting	
	MDI	MDA	MDI	MDA
Flesch Reading Ease	0.1529	0.1046	0.1001	0.1164
Gunning Fog Index	0.1408	0.1239	0.1002	0.2687
SMOG Index	0.1342	0.1374	0.0901	0.3183
LM Index	0.5720	0.9983	0.7095	0.9680

As it can be observed, both machine learning models, Random Forest and Gradient Boosting, either by MDI or MDA feature relevance methods point towards the same result, the LM Index is highly relevant towards predicting the Informativeness Value Score of a financial report. To verify this importance, we use SHAP as an alternative method to interpret how each variable contributes to a prediction model (Futagami et al., 2021). The results are presented in figures 1 and 2

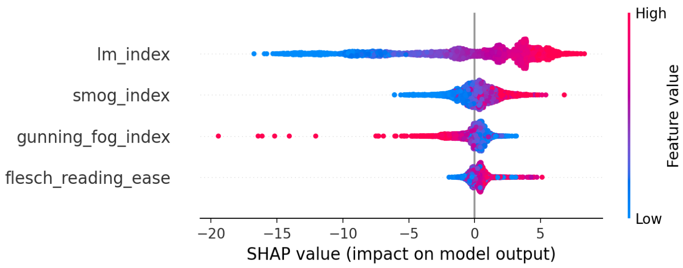


Figure 1 – Gradient Boosting Model SHAP summary

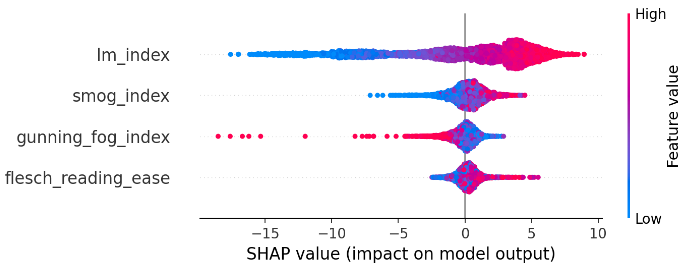


Figure 2 – Random Forest Model SHAP summary

We employ a violin plot to visualize the distribution of SHAP values for each variable. The X-axis displays the SHAP values: higher positive values indicate greater feature relevance, while lower values correspond to lesser importance. The color gradient reflects the magnitude of each feature's value. Analyzing the results for both models, the results provided by the MDI and MDA models are reinforced, as the LM Index is largely the main contributor to the predictability of both models, thus better representing the informational value of financial text.

From an interpretive standpoint, this convergence of evidence across multiple feature importance techniques provides strong support for the conclusion that the LM Index constitutes the most reliable and conceptually coherent readability-related predictor of informativeness in Portuguese financial disclosures due to its approach. The LM Index is calculated based on the size of the file, and as such it carries a higher informational content than the complexity of the text it contains, as a file size is the result of the number of characters in a file (You, 2010). Thereby, by using the file size of a financial report as it readability metric, we are getting a fairly simple quantification of the informational value a file has. While Loughran & McDonald (2014) define readability as the effective communication of valuation-relevant information, the term is usually defined as a measurement of a document's reading difficulty (Brennan et al., 2009; Efretuei & Hussainey, 2023), as such, our findings suggests that, unlike the usually employed readability metrics, the Index proposed by Loughran & McDonald (2014) is one actually capable of serving as proxy for a financial text informativeness.

Our findings challenges the assumption that the ease of reading may be somewhat correlated with quality of the disclosure (Agarwal, 2020). The results, however, suggest a more nuanced relationship: Excessive complexity may have a negative impact on comprehension, nonetheless, a certain level of textual density or volume may lead to better informativeness, especially with highly technical and regulated context, such as financial reporting.

Both models show consistent and moderate predictive accuracy, robust under cross-validation, and convergent across three different interpretability techniques. The dominance of the LM Index across all methods provides a strong empirical foundation for its use as a proxy of informativeness in Portuguese-language corporate texts, opening a pathway for future studies to further integrate computational linguistics and financial accounting research. Additionally, the readability metrics in combination with the BERT-based metric allows for an interesting framework for studying the quality of information shown in financial disclosure, while the LM Index has been shown to be able to proxy the informational value of Portuguese financial text.

4.1.4 Readability metrics and Informativeness

The empirical results presented above can be interpreted within the broader theoretical framework of agency theory and the literature on financial disclosure quality. According to the classical formulation of Jensen & Meckling (1976) and Eisenhardt (1989), the relationship between managers (agents) and investors (principals) is characterized by information asymmetry: managers have superior knowledge of the firm's operations and prospects, while outside investors must rely on disclosed financial information to make decisions. One of the key mechanisms through which agency costs can be mitigated is the transparent disclosure of information (Leuz & Verrecchia, 2000; Morris, 1987). However, as several studies have shown, the disclosure process is not neutral; managers may strategically adjust the content and presentation of information to influence investor perceptions (Courtis, 1995; F. Li, 2008; Bushee et al., 2018).

While previous literature has approached the quantification of the informativeness of financial reports through investor response to the disclosure of the financial report (Agarwal, 2020; Merkley, 2014), our proposal is novel in its leverage of a automated method capable of understanding text and quantifying metrics related to the qualitative characteristics of the accounting information.

However, due to our employed data, we are unable to conduct validity tests on market-related metrics, as most of the companies we observe are not listed. As validation for our Index, we explore the a methods commonly employed on accounting data as related to the information carried by financial text, Readability. Drawing from linguistic theory, we understand that readability measures assess the surface structure of text, the syntactic and lexical complexity that affects how easily information can be processed by readers (Vallduví & Engdahl, 1996), when combined with economic and accounting theoretical frameworks, we understand how they may be employed to measure the attempt to distract users from underlying economic message (Bloomfield, 2008). Therefore, by comparing our Informativeness Index with several readability metrics, we evaluate whether our measure captures deeper informational content beyond the mere textual complexity of financial statements. In this context, the findings of this study provide new insights into how textual characteristics relate to the informativeness of financial statement notes.

To that end, the dominance of the LM Index across all feature importance techniques suggests that longer and more extensive disclosures are positively associated with higher informativeness scores as measured by the FinBERT-PT-BR model. From an agency-theory standpoint, this result may indicate that firms engaging in more comprehensive and voluminous reporting tend to provide more substantive content, therefore reducing information asymmetry. Such disclosures, while potentially more complex, appear to convey richer informational signals that FinBERT-PT-BR interprets as semantically dense and contextually informative.

At the same time, this finding invites reflection on the dual nature of textual length in financial reporting. Prior research has argued that verbosity or repetition can serve as an instrument of obfuscation, a deliberate attempt by management to reduce the accessibility of information (Bloomfield, 2008; F. Li, 2008; Bushee et al., 2018; Carlé et al., 2023). Yet, in the Portuguese-language corporate context examined here, longer notes seem to carry a positive informational weight rather than signaling obfuscation. This may stem from institutional and linguistic particularities of Brazilian financial reporting, in which companies often follow prescriptive disclosure standards that demand detailed narrative explanations of accounting estimates, contingencies, and sustainability-related information. Consequently, greater textual volume may reflect compliance and completeness rather than strategic opacity.

These findings therefore refine the traditional assumption that readability, understood merely as ease of reading, necessarily equates to disclosure quality. In contrast, our results suggest that informativeness is more closely linked to semantic richness and contextual depth, which are better captured by measures such as the LM Index. This observation aligns with recent developments in the disclosure literature emphasizing the multidimensional nature of textual quality — encompassing clarity, completeness, and relevance (Hassan et al., 2019; Du & Yu, 2021). From this perspective, readability metrics such as Flesch-Kincaid or Fog Index remain useful indicators of linguistic simplicity, but they do not fully capture the cognitive and informational substance of financial texts.

The use of a large language model (FinBERT-PT-BR) to quantify informativeness also strengthens this theoretical interpretation. Because FinBERT-PT-BR's embeddings encode semantic relationships rather than surface linguistic patterns, the positive association between the LM Index and the informativeness score indicates that textual expansiveness is accompanied by higher conceptual density. This supports the notion that information quality arises not only from syntactic clarity but also from the semantic granularity of disclosure, a dimension that traditional readability metrics fail to measure.

Overall, the results contribute to the ongoing debate between the "clarity" and "completeness" paradigms in financial communication. While earlier studies rooted in the obfuscation hypothesis tended to equate longer or more complex texts with lower transparency (F. Li, 2008; Courtis, 1995), the findings presented here suggest that, in the Brazilian setting, textual elaboration may enhance rather than hinder informativeness. Consequently, the LM Index emerges not merely as a measure of textual size but as a proxy for informational density, capable of capturing the trade-off between verbosity and substance in corporate reporting.

In essence, the evidence supports an interpretation consistent with agency theory's emphasis on disclosure as a mechanism for reducing information asymmetry, while also providing a nuanced view of how textual features contribute to that objective. By demonstrating that informativeness is semantically rather than syntactically driven, this study extends the theoretical understanding of financial communication in emerging markets. It reinforces the potential of combining natural language processing and accounting research to assess disclosure quality with greater precision.

5 Conclusions

5.1 Conclusions

We conduct an empirical analysis on a corpus of 26,804 quarterly financial reports issued by 1,163 publicly traded Brazilian companies between 2011 and 2023, the largest possible sample given CVM's data availability, to understand how the relationship between the commonly employed readability metrics and the information value of a financial text, as estimated by our model. In addition, two ensemble-based machine learning regressors, Random Forest and Gradient Boosting, were employed due to their robustness in handling non-linear relationships and multicollinearity among textual variables (Breiman, 2001; Friedman, 2001), providing us with regressors able to explore our relationship of interest. Lastly, we measure the relative relevance of each variable (readability metric) on the machine learning regressors, seeking the most relevant metric on predicting informativeness, to that end we conduct a feature importance test based on three methods: Mean Decrease in Impurity (MDI), Mean Decrease in Accuracy (MDA), and SHapley Additive exPlanations (SHAP) (Scornet, 2023; Altmann et al., 2010; Lundberg & Lee, 2017). Each method follows a different methodological approach, and as both MDI and MDA are highly disputed in their validity, the implementation of SHAP should provide us with a robust result.

Our findings consistently indicate that the Loughran–McDonald Index (LM Index) is the most effective metric for capturing the informational dimension of Portuguese financial texts. While traditional readability formulas such as Flesch–Kincaid and Gunning Fog Index are valuable for measuring linguistic simplicity, they primarily reflect surface-level textual accessibility. The LM Index, however, is based on the file size, and as such it implicitly incorporates every single quantifiable aspect of the text presented in the file. As a result, it appears this contributes the metric to be closer to the informativeness measured by the FinBERT model, suggesting that longer and denser financial documents are able to convey greater contextual and explanatory content. Nonetheless, this argument is done at the expense of the capability of the LM Index to be useful as a readability measure in the strict sense, however, it seems to performs remarkably well as a proxy for informativeness.

Viewed through the lens of agency theory, our premise is built upon the idea that financial reporting serves as a mechanism for mitigating information asymmetry between managers and investors (Jensen & Meckling, 1976; Eisenhardt, 1989). Understanding financial reports as such agency-cost mitigating tools, they can be manipulated to convey more or less information, as seen per managerial obfuscation and intentional opacity literature (F. Li, 2008; Courtis, 1995; Bushee et al., 2018). Despite this, our results a different dynamic in the Brazilian setting, where the more textual information a given document has, the higher it's information value. This may be due to linguistic characteristics of Portuguese, IFRS-based financial reporting, or even the overall net-positive effect that verbosity and complexity may have in financial text.

Consequently, this study contributes to the ongoing debate on the contents of textual factor and their impacts on financial communication. Our evidence suggests that informativeness goes beyond the ease of reading, but represents the semantic richness and contextual completeness, metrics not able to be quantified by readability metrics. To that end, our BERT-based Informativeness Index is able to deliver a novel method to empirically quantify financial information. In essence, this research provides a methodological novelty and a conceptual enhancement. Methodologically, we present the viability of using large language models to quantify the informational value of corporate narratives presented in financial reports, expanding the toolkit available to accounting researchers. Conceptually, it expands on the usual employment of readability metrics and their relation with disclosure quality, while also presenting how a readability metric, for Portuguese financial text, is able to be employed as proxy for informativeness, a deeper semantic dimension than usually explored by readability metrics.

Despite or novel approach, we understand some of the research shortcomings. First, as Portuguese is not as broadly used or relevant as English, the language may limit the validity of our research, yet, we believe our empirical evidence on emerging markets is relevant due to it's market size, regional relevance and the corpus. Second, we understand our approach lacks a more substantial construct validity, yet, as previously stated, the majority of the companies studied have market-related data, as such, we conduct our validation on different readability metrics.

We expect future research will be able to further validate and expand these findings in different ways. The exploration of studies on different languages are able to validate the proposed informativeness index, or reveal dynamics different than the explored in Portuguese. Experimental research could also validate the BERT's model ability to measure informational value, increasing the validity of our approach. As last, further analyses on informativeness and firm performance may better link textual informativeness and economic results.

Ultimately, by leveraging machine learning methods and large language models under a natural language processing framework with financial text built upon the interpretative depth of agency theory, this study aids the understanding of how textual features shape the informational landscape of financial reporting. The proposed Informativeness Index provides a empirical proxy for measuring informational quality in Portuguese financial text, but also indicates the role LM Index may have in serving as proxy for such metric, contributing for the research agenda in accounting and finance.

References

Abad-Segura, E., González-Zamar, M.-D., & Squillante, M. (2021). Examining the research on business information-entropy correlation in the accounting process of organizations. *Entropy*, *23*(11), 1493.
Cited on page 22.

Abdel-Khalik, A. R. (1974). The entropy law, accounting data, and relevance to decision-making. *The Accounting Review*, *49*(2), 271–283.
Cited on page 22.

Adler, A. I., & Painsky, A. (2022). Feature importance in gradient boosting trees with cross-validation feature selection. *Entropy*, *24*(5), 687.
Cited 2 times on pages 26 and 28.

Agarwal, N. (2020). Integrated reporting and the informativeness of annual reports. does textual coherence matter? *Does textual coherence matter*.
Cited 3 times on pages 15, 36, and 37.

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, *9*(8), 1295.
Cited on page 22.

Ahn, M., Jung, D., Kim, J.-T., Lee, W.-J., & Sunwoo, H.-Y. (2023). Do more readable sustainability reports provide more value-relevant information to shareholders? *Finance Research Letters*, *57*, 104154.
Cited on page 20.

Akerlof, G. A. (1978). The market for “lemons”: Quality uncertainty and the market mechanism. In *Uncertainty in economics* (pp. 235–251). Elsevier.
Cited on page 16.

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347.
Cited 2 times on pages 34 and 39.

A. Ramezan, C., A. Warner, T., & E. Maxwell, A. (2019). Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification. *Remote Sensing*, *11*(2), 185.
Cited on page 33.

Bag, S., Kumar, S. K., & Tiwari, M. K. (2019). An efficient recommendation generation using relevant jaccard similarity. *Information Sciences*, *483*, 53–64.
Cited on page 21.

Ball, R., & Brown, P. (2013). An empirical evaluation of accounting income numbers. In *Financial accounting and equity markets* (pp. 27–46). Routledge.
Cited on page 20.

Barnett, A., & Leoffler, K. (1979). Readability of accounting and auditing messages. *The Journal of Business Communication (1973)*, *16*(3), 49–59.
Cited on page 18.

Beaver, W. H. (1968). The information content of annual earnings announcements. *Journal of Accounting Research*, *6*, 67–92.
Cited 2 times on pages 13 and 20.

Bengio, Y. (2000, August). Gradient-based optimization of hyperparameters. *Neural Computation*, *12*(8), 1889–1900. Retrieved from `http://dx.doi.org/10.1162/089976600300015187` doi: 10.1162/089976600300015187
Cited on page 33.

Bengio, Y., Goodfellow, I., & Courville, A. (2017). *Deep learning* (Vol. 1). MIT press Cambridge, MA, USA.
Cited on page 32.

Bholowalia, P., & Kumar, A. (2014). Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, *105*(9).
Cited on page 22.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.
Cited on page 22.

Bloomfield, R. (2008). Discussion of "annual report readability, current earnings, and earnings persistence". *Journal of Accounting and Economics*, *45*(2-3), 248–252.
Cited on page 37.

Bochkay, K., Brown, S. V., Leone, A. J., & Tucker, J. W. (2023). Textual analysis in accounting: What's next? *Contemporary accounting research*, *40*(2), 765–805.
Cited 5 times on pages 13, 21, 24, 25, and 27.

Bonsall IV, S. B., Leone, A. J., Miller, B. P., & Rennekamp, K. (2017). A plain english measure of financial reporting readability. *Journal of Accounting and Economics*, *63*(2-3), 329–357.

Cited 2 times on pages 17 and 19.

Bradbury, M. E., Hsiao, P. K., & Scott, T. (2018). Summary annual reports: length, readability and content. *Accounting & Finance*, *60*(3), 2145–2165.
Cited 2 times on pages 13 and 17.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. doi: 10.1023/a:1010933404324
Cited on page 39.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification and regression trees*. Routledge. doi: 10.1201/9781315139470
Cited on page 25.

Brennan, N. M., Guillamon-Saorin, E., & Pierce, A. (2009). Methodological insights: Impression management: developing and illustrating a scheme of analysis for narrative disclosures–a methodological note. *Accounting, Auditing & Accountability Journal*, *22*(5), 789–832.
Cited on page 36.

Brown, S. V., Ma, G., & Tucker, J. W. (2023). Financial statement similarity. *Contemporary Accounting Research*, *40*(4), 2577–2615.
Cited on page 21.

Bushee, B. J., Gow, I. D., & Taylor, D. J. (2018). Linguistic complexity in firm disclosures: Obfuscation or information? *Journal of Accounting Research*, *56*(1), 85–121.
Cited 4 times on pages 13, 21, 37, and 39.

Cameron, A. C., & Windmeijer, F. A. (1997). An r-squared measure of goodness of fit for some common nonlinear regression models. *Journal of econometrics*, *77*(2), 329–342.
Cited on page 32.

Cao, S., Jiang, W., Yang, B., & Zhang, A. L. (2023). How to talk when a machine is listening: Corporate disclosure in the age of ai. *The Review of Financial Studies*, *36*(9), 3603–3642.
Cited 2 times on pages 13 and 15.

Carlé, T., Pappert, N., & Quick, R. (2023). Text similarity, boilerplates and their determinants in key audit matters disclosure. *Corporate Ownership and Control*, *20*(2).
Cited 2 times on pages 21 and 37.

Chen, T.-K., & Tseng, Y. (2021). Readability of notes to consolidated financial statements and corporate bond yield spread. *European Accounting Review*, *30*(1), 83–113.
Cited 4 times on pages 17, 18, 19, and 20.

Cheung, A., & Hu, W. (2019). Information disclosure quality: Correlation versus precision. *Accounting & Finance*, *59*(2), 1033–1053.
Cited 2 times on pages 13 and 17.

Cheung, E., & Lau, J. (2016). Readability of notes to the financial statements and the adoption of ifrs. *Australian Accounting Review*, *26*(2), 162–176.
Cited 3 times on pages 13, 16, and 17.

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *Peerj computer science*, *7*, e623.
Cited on page 32.

Chowdhary, K. R. (2020). Natural language processing. In *Fundamentals of artificial intelligence* (p. 603–649). Springer India. doi: 10.1007/978-81-322-3972-7_19
Cited on page 24.

Clatworthy, M., & Jones, M. J. (2001). The effect of thematic structure on the variability of annual report readability. *Accounting, Auditing & Accountability Journal*, *14*(3), 311–326.
Cited 2 times on pages 13 and 17.

Courtis, J. K. (1995). Readability of annual reports: Western versus asian evidence. *Accounting, Auditing & Accountability Journal*, *8*(2), 4–17.
Cited 4 times on pages 16, 37, 38, and 39.

Courtis, J. K. (1998). Annual report readability variability: tests of the obfuscation hypothesis. *Accounting, Auditing & Accountability Journal*, *11*(4), 459–472.
Cited on page 18.

Courtis, J. K. (2004). Corporate report obfuscation: artefact or phenomenon? *The British Accounting Review*, *36*(3), 291–312.
Cited 2 times on pages 16 and 18.

Datta, A., Sen, S., & Zick, Y. (2016). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 ieee symposium on security and privacy (sp)* (pp. 598–617).
Cited on page 26.

Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
Cited on page 26.

Disha, R. A., & Waheed, S. (2022). Performance analysis of machine learning models for intrusion detection system using gini impurity-based weighted random forest (giwrf) feature selection technique. *Cybersecurity*, *5*(1), 1.
Cited on page 34.

Doyle, G., & Elkan, C. (2009). Accounting for burstiness in topic models. In *Proceedings of the 26th annual international conference on machine learning* (pp. 281–288).
Cited on page 22.

Du, S., & Yu, K. (2021). Do corporate social responsibility reports convey value relevant information? evidence from report readability and tone. *Journal of business ethics*, *172*(2), 253–274.
Cited 2 times on pages 17 and 38.

Dyer, T., Lang, M., & Stice-Lawrence, L. (2017). The evolution of 10-k textual disclosure: Evidence from latent dirichlet allocation. *Journal of Accounting and Economics*, *64*(2-3), 221–245.
Cited 2 times on pages 13 and 17.

Edossa, F. W., Gassen, J., & Maas, V. S. (2024). Using large language models to explore contextualization effects in economics-based accounting experiments. *Available at SSRN 4891763*.
Cited on page 27.

Efretuei, E., & Hussainey, K. (2023). The fog index in accounting research: contributions and challenges. *Journal of Applied Accounting Research*, *24*(2), 318–343.
Cited on page 36.

Efretuei, E., Usoro, A., & Koutra, C. (2022). Complex information and accounting standards: Evidence from uk narrative reporting. *South African Journal of Accounting Research*, *36*(3), 171–194.
Cited 3 times on pages 13, 18, and 19.

Eisenhardt, K. M. (1989). Agency theory: An assessment and review. *Academy of management review*, *14*(1), 57–74.
Cited 4 times on pages 13, 16, 36, and 39.

El-Haj, M., Alves, P., Rayson, P., Walker, M., & Young, S. (2020). Retrieving, classifying and analysing narrative commentary in unstructured (glossy) annual reports published as pdf files. *Accounting and Business Research*, *50*(1), 6–34.
Cited on page 29.

Engel, C., Grossmann, M. R., & Ockenfels, A. (2024). Integrating machine behavior into human subject experiments: A user-friendly toolkit and illustrations. *MPI Collective Goods Discussion Paper*(2024/1).
Cited on page 27.

Ferreira, A. J., & Figueiredo, M. A. (2012). Boosting algorithms: A review of methods, theory, and applications. *Ensemble machine learning: Methods and applications*, 35–85.
Cited on page 25.

Ferri, P., Lusiani, M., & Pareschi, L. (2021). Shades of theory: A topic modelling of ways of theorizing in accounting history research. *Accounting History*, *26*(3), 484–519.
Cited on page 22.

Fiordelisi, F., & Ricci, O. (2014). Corporate culture and ceo turnover. *Journal of Corporate Finance*, *28*, 66–82.
Cited on page 13.

Fisher, I. E., Garnsey, M. R., & Hughes, M. E. (2016). Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. *Intelligent Systems in Accounting, Finance and Management*, *23*(3), 157–214.
Cited on page 24.

Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, *32*(3), 221.
Cited on page 18.

Fontes, A., Rodrigues, L. L., & Craig, R. (2005). Measuring convergence of national accounting standards with international financial reporting standards. In *Accounting forum* (Vol. 29, pp. 415–436).
Cited on page 21.

Foundation, I. (2018). Conceptual framework for financial reporting. *IFRS Foundation*.
Cited 3 times on pages 16, 17, and 20.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, *55*(1), 119–139.
Cited on page 25.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.
Cited 2 times on pages 25 and 39.

Friedman, J. H. (2002, feb). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378. doi: 10.1016/s0167-9473(01)00065-2

Cited 2 times on pages 13 and 25.

Futagami, K., Fukazawa, Y., Kapoor, N., & Kito, T. (2021). Pairwise acquisition prediction with shap value interpretation. *The Journal of Finance and Data Science*, *7*, 22–44.
Cited on page 35.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.
Cited on page 22.

Gunawan, D., Sembiring, C., & Budiman, M. A. (2018). The implementation of cosine similarity to calculate text relevance between two documents. In *Journal of physics: conference series* (Vol. 978, p. 012120).
Cited on page 21.

Gunning, R. (1952). *The technique of clear writing*. McGraw-Hill.
Cited on page 19.

Guo, K. (2022). Testing and validating the cosine similarity measure for textual analysis. *Available at SSRN 4258463*.
Cited on page 21.

Hall, M. A. (1999). *Correlation-based feature selection for machine learning* (PhD thesis). The University of Waikato.
Cited on page 34.

Hall, M. A. (2000). *Correlation-based feature selection of discrete and numeric class machine learning* (PhD thesis). The University of Waikato.
Cited on page 34.

Hassan, M. K., Abu Abbas, B., & Garas, S. N. (2019). Readability, governance and performance: a test of the obfuscation hypothesis in qatari listed firms. *Corporate Governance: The International Journal of Business in Society*, *19*(2), 270–298.
Cited 3 times on pages 16, 17, and 38.

Hemmings, D., Hodgkinson, L., & Williams, G. (2020). It's ok to pay well, if you write well: The effects of remuneration disclosure readability. *Journal of Business Finance & Accounting*, *47*(5-6), 547–586.
Cited on page 19.

Hindman, M. (2015). Building better models: Prediction, replication, and machine learning in the social sciences. *The Annals of the American Academy of Political and Social Science*, *659*(1), 48–62.
Cited on page 25.

Hoberg, G., & Lewis, C. (2017). Do fraudulent firms produce abnormal disclosure? *Journal of Corporate Finance*, *43*, 58–85.
Cited on page 18.

Holthausen, R. W., & Leftwich, R. W. (1983). The economic consequences of accounting choice implications of costly contracting and monitoring. *Journal of accounting and economics*, *5*, 77–117.
Cited on page 16.

Holthausen, R. W., & Watts, R. L. (2001). The relevance of the value-relevance literature for financial accounting standard setting. *Journal of accounting and economics*, *31*(1-3), 3–75.
Cited on page 20.

Huang, A. H., Wang, H., & Yang, Y. (2023). Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, *40*(2), 806–841.
Cited on page 27.

Jabarian, B., & Imas, A. (2025). *Artificial writing and automated detection* (Tech. Rep.). National Bureau of Economic Research.
Cited on page 15.

Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, *3*(4), 305-360.
Cited 4 times on pages 16, 17, 36, and 39.

Ji, J., Li, J., Yan, S., Tian, Q., & Zhang, B. (2013). Min-max hash for jaccard similarity. In *2013 ieee 13th international conference on data mining* (pp. 301–309).
Cited on page 21.

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: A developmental perspective. *Working papers/Lund University, Department of Linguistics and Phonetics*, *53*, 61–79.
Cited on page 23.

Johnston, J. A., & Zhang, J. H. (2021). Auditor style and financial reporting similarity. *Journal of Information Systems*, *35*(1), 79–99.
Cited on page 21.

Jones, M., & Smith, M. (2014). Traditional and alternative methods of measuring the understandability of accounting narratives. *Accounting, Auditing & Accountability Journal*, *27*(1), 183–208.
Cited on page 16.

Jones, M. J., & Shoemaker, P. A. (1994). Accounting narratives: A review of empirical studies of content and readability. *Journal of accounting literature*, *13*, 142.
Cited on page 17.

Jurafsky, D., & Martin, J. H. (2024). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models* (3rd ed.).
Cited on page 26.

Kim, Y., & Kim, Y. (2022). Explainable heat-related mortality with random forest and shapley additive explanations (shap) models. *Sustainable Cities and Society*, *79*, 103677.
Cited on page 28.

Krishna, K., & Murty, M. N. (1999). Genetic k-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *29*(3), 433–439.
Cited on page 22.

Lahitani, A. R., Permanasari, A. E., & Setiawan, N. A. (2016). Cosine similarity to determine similarity measure: Study case in online essay assessment. In *2016 4th international conference on cyber and it service management* (pp. 1–6).
Cited on page 21.

Lang, M., & Stice-Lawrence, L. (2015). Textual analysis and international financial reporting: Large sample evidence. *Journal of Accounting and Economics*, *60*(2-3), 110–135.
Cited 2 times on pages 19 and 20.

Lee, C. S., Cheang, P. Y. S., & Moslehpour, M. (2022). Predictive analytics in business analytics: decision tree. *Advances in Decision Sciences*, *26*(1), 1–29.
Cited on page 25.

Leuz, C. (2010). Different approaches to corporate reporting regulation: How jurisdictions differ and why. *Accounting and business research*, *40*(3), 229–256.
Cited on page 14.

Leuz, C., & Verrecchia, R. E. (2000). The economic consequences of increased disclosure. *Journal of accounting research*, 91–124.
Cited 2 times on pages 16 and 36.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and economics*, *45*(2-3), 221–247.
Cited 6 times on pages 17, 18, 19, 37, 38, and 39.

Li, K., Mai, F., Shen, R., & Yan, X. (2021). Measuring corporate culture using machine learning. *The Review of Financial Studies*, *34*(7), 3265–3315.
Cited 3 times on pages 13, 24, and 26.

Liang, J., Shi, Z., Li, D., & Wierman, M. J. (2006). Information entropy, rough entropy and knowledge granulation in incomplete information systems. *International Journal of general systems*, *35*(6), 641–654.
Cited on page 22.

Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, *36*(2), 451–461.
Cited on page 22.

Linsley, P. M., & Lawrence, M. J. (2007). Risk reporting by the largest uk companies: readability and lack of obfuscation. *Accounting, Auditing & Accountability Journal*, *20*(4), 620–627.
Cited on page 13.

Loughran, T., & McDonald, B. (2014). Measuring readability in financial disclosures. *the Journal of Finance*, *69*(4), 1643–1671.
Cited 4 times on pages 17, 18, 19, and 36.

Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, *54*(4), 1187–1230.
Cited 2 times on pages 19 and 26.

Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. (2013). Understanding variable importances in forests of randomized trees. *Advances in neural information processing systems*, *26*.
Cited 2 times on pages 26 and 28.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, *2*(1), 56–67.
Cited on page 34.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.
Cited 3 times on pages 26, 34, and 39.

Lundholm, R. J., Rogo, R., & Zhang, J. L. (2014). Restoring the tower of babel: How foreign firms communicate with us investors. *The accounting review*, *89*(4), 1453–1485.
Cited on page 19.

Martins, T. B., Ghiraldelo, C. M., Nunes, M. d. G. V., & Oliveira Junior, O. N. d. (1996). *Readability formulas applied to textbooks in brazilian portuguese* (Research paper). Universidade de São Paulo.
Cited on page 18.

Mayr, A., Binder, H., Gefeller, O., & Schmid, M. (2014). The evolution of boosting algorithms. *Methods of information in medicine*, *53*(06), 419–427.
Cited on page 25.

Mc Laughlin, G. H. (1969). Smog grading - a new readability formula. *Journal of reading*, *12*(8), 639–646.
Cited on page 19.

Meade, N. (1983). *Industrial and business forecasting methods, lewis, cd, borough green, sevenoaks, kent: Butterworth, 1982. price:£ 9.25. pages: 144.* Wiley Online Library.
Cited on page 32.

Merkley, K. J. (2014). Narrative disclosure and earnings performance: Evidence from r&d disclosures. *The Accounting Review*, *89*(2), 725–757.
Cited 2 times on pages 15 and 37.

Michelon, G., Sealy, R., & Trojanowski, G. (2020). *Understanding research findings and evidence on corporate reporting: An independent literature review* (Tech. Rep.). Financial Reporting Council.
Cited on page 13.

Mikolov, T. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
Cited on page 26.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.
Cited on page 26.

Moreno, J. J. M., Pol, A. P., Abad, A. S., & Blasco, B. C. (2013). Using the r-mape index as a resistant measure of forecast accuracy. *Psicothema*, *25*(4), 500–506.
Cited on page 32.

Morris, R. D. (1987). Signalling, agency theory and accounting policy choice. *Accounting and business Research*, *18*(69), 47–56.
Cited 2 times on pages 16 and 36.

Mosca, E., Szigeti, F., Tragianni, S., Gallagher, D., & Groh, G. (2022). Shap-based explanation methods: a review for nlp interpretability. In *Proceedings of the 29th international conference on computational linguistics* (pp. 4593–4603).
Cited 3 times on pages 26, 28, and 34.

Nadeem, M. (2021, May). Board gender diversity and managerial obfuscation: Evidence from the readability of narrative disclosure in 10-k reports. *Journal of Business Ethics*, *179*(1), 153–177. Retrieved from http://dx.doi.org/10.1007/s10551-021-04830-3 doi: 10.1007/s10551-021-04830-3
Cited on page 13.

Nielsen, S. (2022). Management accounting and the concepts of exploratory data analysis and unsupervised machine learning: a literature study and future directions. *Journal of Accounting & Organizational Change*, *18*(5), 811–853.
Cited on page 28.

Pavlyshenko, B. (2018). Using stacking approaches for machine learning models. In *2018 ieee second international conference on data stream mining & processing (dsmp)* (pp. 255–258).
Cited on page 28.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). *Deep contextualized word representations.* arXiv. doi: 10.48550/ARXIV.1802.05365
Cited on page 26.

Pishro-Nik, H. (2014). *Introduction to probability, statistics, and random processes.* Kappa Research.
Cited on page 32.

Quinlan, J. R. (1996). Learning decision tree classifiers. *ACM Computing Surveys (CSUR)*, *28*(1), 71–72.
Cited on page 25.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). *Improving language understanding by generative pre-training* (Technical report). OpenAI.
Cited on page 26.

Ranta, M., Ylinen, M., & Järvenpää, M. (2023). Machine learning in management accounting research: Literature review and pathways for the future. *European Accounting Review*, *32*(3), 607–636.
Cited 3 times on pages 14, 25, and 28.

Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the fourth berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics* (Vol. 4, pp. 547–562).
Cited on page 22.

Roth, A. E. (1988). *The shapley value: essays in honor of lloyd s. shapley*. Cambridge University Press.
Cited on page 26.

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, *8*(4), e1249.
Cited on page 24.

Salehi, M., Lari Dasht Bayaz, M., Mohammadi, S., Adibian, M. S., & Fahimifard, S. H. (2020). Auditors' response to readability of financial statement notes. *Asian Review of Accounting*, *28*(3), 463–480.
Cited 2 times on pages 17 and 18.

Santos, L. L., Bianchi, R. A., & Costa, A. H. (2023). Finbert-pt-br: Análise de sentimentos de textos em português do mercado financeiro. In *Anais do ii brazilian workshop on artificial intelligence in finance* (pp. 144–155).
Cited 2 times on pages 14 and 27.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, *2*(3), 160.
Cited 2 times on pages 14 and 25.

Sarker, I. H., Furhad, M. H., & Nowrozy, R. (2021). Ai-driven cybersecurity: an overview, security intelligence modeling and research directions. *SN Computer Science*, *2*(3), 173.
Cited on page 14.

Schapire, R. E. (2003). The boosting approach to machine learning: An overview. *Nonlinear estimation and classification*, 149–171.
Cited on page 25.

Schapire, R. E., & Freund, Y. (2013). Boosting: Foundations and algorithms. *Kybernetes*, *42*(1), 164–166.
Cited on page 25.

Schomacker, T., & Tropmann-Frick, M. (2021). Language representation models: An overview. *Entropy*, *23*(11), 1422.
Cited on page 26.

Schütze, H., Manning, C. D., & Raghavan, P. (2008). *Introduction to information retrieval* (Vol. 39). Cambridge University Press Cambridge.
Cited on page 21.

Scornet, E. (2023). Trees, forests, and impurity-based variable importance in regression. In *Annales de l'institut henri poincare (b) probabilites et statistiques* (Vol. 59, pp. 21–52).
Cited 3 times on pages 26, 34, and 39.

SEC. (2013). *Report on review of disclosure requirements in regulation s-k* (Report). U.S. Securities and Exchange Commission.
Cited on page 17.

Senave, E., Jans, M. J., & Srivastava, R. P. (2023). The application of text mining in accounting. *International Journal of Accounting Information Systems*, *50*, 100624.
Cited on page 13.

Shahapure, K. R., & Nicholas, C. (2020). Cluster quality analysis using silhouette score. In *2020 ieee 7th international conference on data science and advanced analytics (dsaa)* (pp. 747–748).
Cited on page 22.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379–423.
Cited on page 22.

Silva, C. A. T., & Fernandes, J. L. T. (2009). Legibilidade dos fatos relevantes no brasil. *RAC-eletrônica*, *3*(1), 142–158.
Cited on page 18.

Smith, J. E., & Smith, N. P. (1971). Readability: A measure of the performance of the communication function of financial reporting. *The Accounting Review*, *46*(3), 552–561.
Cited 2 times on pages 17 and 18.

Song, Y.-Y., & Ying, L. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, *27*(2), 130.
Cited on page 25.

Srivastava, R. P. (2023). A new measure of similarity in textual analysis: Vector similarity metric versus cosine similarity metric. *Journal of Emerging Technologies in Accounting*, *20*(1), 77–90.
Cited on page 21.

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, *41*(3), 647–665.
Cited on page 26.

Syakur, M. A., Khotimah, B. K., Rochman, E. M., & Satoto, B. D. (2018). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In *Iop conference series: materials science and engineering* (Vol. 336, p. 012017).
Cited on page 22.

Tan, Z., Yan, Z., & Zhu, G. (2019). Stock selection with random forest: An exploitation of excess return in the chinese stock market. *Heliyon*, *5*(8).
Cited on page 28.

Telles, S. V., & Salotti, B. M. (2024). Readability and understandability of notes to financial statements. *Revista Brasileira de Gestão de Negócios*, *26*, e20230127.
Cited on page 20.

Thiprungsri, S., & Vasarhelyi, M. A. (2011). Cluster analysis for anomaly detection in accounting data: An audit approach. *International Journal of Digital Accounting Research*, *11*.
Cited on page 22.

Tougui, I., Jilbab, A., & El Mhamdi, J. (2021). Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications. *Healthcare informatics research*, *27*(3), 189–199.
Cited on page 33.

Travieso, G., Benatti, A., & Costa, L. d. F. (2024). An analytical approach to the jaccard similarity index. *arXiv preprint arXiv:2410.16436*.
Cited on page 21.

Tung, W., Quek, C., & Cheng, P. (2004). Genso-ews: a novel neural-fuzzy based early warning system for predicting bank failures. *Neural networks*, *17*(4), 567–587.
Cited on page 24.

Vallduví, E., & Engdahl, E. (1996). The linguistic realization of information packaging. *Linguistics*, *34*(3), 459–520.
Cited on page 37.

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.
Cited on page 26.

Vergani, A. A., & Binaghi, E. (2018). A soft davies-bouldin separation measure. In *2018 ieee international conference on fuzzy systems (fuzz-ieee)* (pp. 1–8).
Cited on page 22.

Wang, D.-n., Li, L., & Zhao, D. (2022). Corporate finance risk prediction based on lightgbm. *Information Sciences*, *602*, 259–268.
Cited 2 times on pages 25 and 28.

Wang, X., & Xu, Y. (2019). An improved index for clustering validation based on silhouette index and calinski-harabasz index. In *Iop conference series: Materials science and engineering* (Vol. 569, p. 052024).
Cited on page 22.

Xia, P., Zhang, L., & Li, F. (2015). Learning similarity with cosine similarity ensemble. *Information sciences*, *307*, 39–52.
Cited on page 21.

Yang, M. (2024). Topic modeling of financial accounting research over 70 years. *International Studies of Economics*, *19*(4), 617–643.
Cited on page 22.

You, Y. (2010). *Audio coding: theory and applications*. Springer Science & Business Media.
Cited on page 36.

Zou, Z. B., Peng, H., & Luo, L. K. (2015). The application of random forest in finance. *Applied Mechanics and Materials*, *740*, 947–951.
Cited on page 28.

Appendix

APPENDIX A – Additional Data Tables

Table 7 – Summary - Financial Reports per Sector per Year

Sector	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Agriculture	33	33	33	33	27	30	30	30	30	39	41	39	39
Banking	94	98	96	99	98	93	90	99	92	90	96	99	90
Civil Construction	116	110	111	110	108	108	104	99	101	134	143	140	145
Commerce	63	63	66	71	68	66	71	72	74	119	149	145	146
Communication and Informatics	17	21	21	21	24	21	21	18	24	33	55	54	53
Drinks and Tabacco	8	6	9	6	6	3	3	3	3	3	3	6	6
Education	6	6	12	12	12	12	12	15	15	18	18	21	24
Electricity	168	169	167	169	169	162	168	177	183	201	226	237	237
Factoring	1	0	0	0	0	0	0	0	0	0	0	0	0
Financial Intermediary	9	9	9	9	9	9	7	9	6	6	12	12	9
Food	42	36	34	34	33	29	33	30	30	30	39	36	39
Graphical Design and Publishing	3	3	3	3	3	3	3	3	1	0	0	0	0
Hospitality and Tourism	24	24	25	24	24	18	18	18	17	15	15	17	16
Insurance	9	9	9	9	12	12	15	15	15	15	15	15	15
Leasing	33	31	30	30	29	27	25	24	20	16	11	9	9
Machines, Equipment, Vehicles and Parts	88	88	87	86	81	81	81	79	76	77	83	81	84
Medical Services	12	12	10	9	9	15	18	21	24	30	48	48	45
Metallurgy and Steel	63	54	51	51	48	48	48	48	45	45	45	45	45
Mineral Extraction	9	9	9	9	9	9	9	9	9	15	16	15	15
Oil and Gas	12	10	9	6	6	6	6	6	6	9	9	9	18
Packaging	10	9	9	9	9	9	9	9	9	9	9	9	9
Petrochemicals and Rubber	35	33	33	33	33	28	27	27	27	25	27	30	30
Pharmaceuticals and Hygiene	21	24	24	21	21	21	21	24	25	21	30	33	33
Pulp and Paper	18	21	21	21	21	21	18	18	15	15	12	12	12
Real Estate Credit	3	4	3	3	3	3	3	3	3	3	3	3	3
Reforestation	7	3	3	3	3	3	3	3	3	3	3	3	3
Sanitization and Utilities	40	39	42	42	45	45	45	45	48	54	58	69	72
Securities	154	161	174	181	193	175	177	180	184	182	201	166	82
Stock Exchange	3	3	3	3	3	3	3	3	3	3	3	3	3
Telecommuncations	46	32	27	27	21	21	21	18	18	18	24	30	36
Textile Industries	83	84	81	71	69	66	63	65	60	66	67	63	63
Toys and Leisure	15	16	15	15	15	18	18	18	18	18	18	18	17
Transport and Logistics	154	157	156	171	191	198	200	190	197	212	225	237	262
Total	**1399**	**1377**	**1382**	**1391**	**1402**	**1363**	**1370**	**1378**	**1381**	**1524**	**1704**	**1704**	**1660**

Table 8 – Summary - Financial Reports per Sector per Year (Holding Companies)

Sector	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Holding Company - Agriculture	3	3	3	3	3	3	3	3	3	6	6	6	3
Holding Company - Civil Construction	39	35	33	33	33	32	33	32	30	39	43	40	39
Holding Company - Commerce	27	28	27	25	24	24	27	27	20	21	25	27	30
Holding Company - Communication and Informatics	0	0	3	3	3	3	3	3	3	6	6	9	12
Holding Company - Education	12	15	15	11	9	9	9	9	9	9	9	9	9
Holding Company - Electricity	65	67	66	69	68	65	66	72	66	72	86	91	88
Holding Company - Financial Intermediary	15	18	18	18	18	15	15	15	12	9	12	18	15
Holding Company - Food	12	15	13	12	12	10	9	9	9	9	9	9	9
Holding Company - Graphical Design and Publishing	3	3	3	3	3	3	0	0	0	0	0	0	0
Holding Company - Hospitality and Tourism	8	7	6	6	4	3	6	6	6	6	6	3	3
Holding Company - Insurance	9	9	9	9	11	12	12	12	12	12	12	12	9
Holding Company - Leasing	6	5	3	3	3	3	0	0	0	0	0	0	0
Holding Company - Machines, Equipment, Vehicles and Parts	17	15	15	15	12	12	15	18	18	18	20	18	18
Holding Company - Medical Services	3	3	3	3	3	0	0	3	6	6	9	6	1
Holding Company - Metallurgy and Steel	18	18	15	15	15	15	15	15	15	15	15	15	15
Holding Company - Mineral Extraction	15	21	20	15	15	13	15	15	12	12	12	12	9
Holding Company - No Main Sector	177	193	178	154	159	148	143	136	118	99	120	113	111
Holding Company - Oil and Gas	9	8	6	6	12	14	12	15	15	18	24	28	30
Holding Company - Petrochemicals and Rubber	9	7	6	5	3	3	3	3	3	3	3	3	2
Holding Company - Pharmaceuticals and Hygiene	3	3	3	0	0	0	3	3	3	2	0	0	0
Holding Company - Pulp and Paper	3	3	3	3	3	3	3	3	3	3	3	3	3
Holding Company - Real Estate Credit	12	12	12	12	12	12	10	9	9	9	9	9	9
Holding Company - Sanitization and Utilities	6	6	6	9	6	6	6	6	8	12	12	15	15
Holding Company - Securities	12	10	6	6	6	6	6	6	6	6	6	6	4
Holding Company - Stock Exchange	1	0	0	0	0	0	0	0	0	0	0	0	0
Holding Company - Telecommunications	54	54	51	51	41	34	29	23	15	14	18	18	18
Holding Company - Textile Industries	9	7	6	6	6	6	6	6	6	6	9	6	6
Holding Company - Toys and Leisure	3	3	3	3	3	3	3	3	3	3	6	6	3
Holding Company - Transport and Logistics	45	47	48	38	36	36	39	39	41	37	41	39	39
Total	**595**	**615**	**580**	**536**	**523**	**493**	**491**	**491**	**451**	**452**	**521**	**521**	**500**

Table 9 – Summary - Reporting companies per Sector per Year

Sector	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Agriculture	12	11	12	11	9	10	10	10	10	13	14	13	13
Banking	32	34	32	33	33	31	30	33	31	30	32	33	31
Civil Construction	39	37	37	37	36	37	35	33	35	45	48	47	49
Commerce	21	21	22	24	23	22	24	24	25	40	51	49	49
Communication and Informatics	6	7	7	7	8	7	7	6	8	11	19	18	18
Drinks and Tabacco	3	2	3	2	2	1	1	1	1	1	1	2	2
Education	2	2	4	4	4	4	4	5	5	6	6	7	8
Electricity	56	56	56	57	57	55	56	59	61	67	75	79	79
Factoring	1	0	0	0	0	0	0	0	0	0	0	0	0
Financial Intermediary	3	3	3	3	3	3	3	3	2	2	4	4	3
Food	14	12	11	12	11	11	11	10	10	10	13	12	13
Graphical Design and Publishing	1	1	1	1	1	1	1	1	1	0	0	0	0
Hospitality and Tourism	8	8	9	8	8	6	6	6	6	5	5	6	6
Insurance	3	3	3	3	4	4	5	5	5	5	5	5	5
Leasing	11	11	10	10	10	9	9	8	7	6	4	3	3
Machines, Equipment, Vehicles and Parts	30	29	29	29	27	27	27	27	26	26	28	27	28
Medical Services	4	4	4	3	3	5	6	7	8	10	16	16	16
Metallurgy and Steel	21	19	17	17	16	16	16	16	15	15	15	15	15
Mineral Extraction	3	3	3	3	3	3	3	3	3	5	6	5	5
Oil and Gas	5	4	3	2	2	2	2	2	2	3	3	3	6
Packaging	4	3	3	3	3	3	3	3	3	3	3	3	3
Petrochemicals and Rubber	12	11	11	11	11	10	9	9	9	9	9	10	10
Pharmaceuticals and Hygiene	7	8	8	7	7	7	7	8	9	7	10	11	11
Pulp and Paper	6	7	7	7	7	7	6	6	5	5	4	4	4
Real Estate Credit	1	1	1	1	1	1	1	1	1	1	1	1	1
Reforestation	3	1	1	1	1	1	1	1	1	1	1	1	1
Sanitization and Utilities	14	13	14	14	15	15	15	15	16	18	19	23	24
Securities	52	55	58	61	67	60	60	65	62	64	70	74	28
Stock Exchange	1	1	1	1	1	1	1	1	1	1	1	1	1
Telecommuncations	16	11	9	9	7	7	7	6	6	6	8	10	12
Textile Industries	28	28	28	24	23	22	21	22	20	22	22	21	21
Toys and Leisure	5	6	5	5	5	6	6	6	6	6	6	6	6
Transport and Logistics	52	53	52	57	64	67	67	64	66	71	75	79	88
Total	**476**	**465**	**464**	**467**	**472**	**461**	**460**	**466**	**466**	**514**	**574**	**588**	**559**

Table 10 – Summary - Reporting companies per Sector per Year (Holding Companies)

Sector	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Holding Company - Agriculture	1	1	1	1	1	1	1	1	1	2	2	2	1
Holding Company - Civil Construction	13	13	11	11	11	11	11	12	10	13	14	14	13
Holding Company - Commerce	9	9	9	9	8	8	9	9	7	7	9	9	10
Holding Company - Communication and Informatics	0	0	1	1	1	1	1	1	1	2	2	3	4
Holding Company - Education	4	5	5	4	3	3	3	3	3	3	3	3	3
Holding Company - Electricity	22	23	22	23	23	22	22	25	22	24	29	30	30
Holding Company - Financial Intermediary	5	6	6	6	6	5	5	5	5	3	4	6	5
Holding Company - Food	4	5	5	4	4	4	3	3	3	3	3	3	3
Holding Company - Graphical Design and Publishing	1	1	1	1	1	1	0	0	0	0	0	0	0
Holding Company - Hospitality and Tourism	3	3	2	2	2	1	2	2	2	2	2	1	1
Holding Company - Insurance	3	3	3	3	4	4	4	4	4	4	4	4	3
Holding Company - Leasing	2	2	1	1	1	1	0	0	0	0	0	0	0
Holding Company - Machines, Equipment, Vehicles and Parts	6	5	5	5	4	4	5	6	6	6	7	6	6
Holding Company - Medical Services	1	1	1	1	1	0	0	1	2	2	3	2	1
Holding Company - Metallurgy and Steel	6	6	5	5	5	5	5	5	5	5	5	5	5
Holding Company - Mineral Extraction	5	7	7	5	5	5	5	5	4	4	4	4	3
Holding Company - No Main Sector	61	67	60	53	55	52	48	46	41	33	41	39	37
Holding Company - Oil and Gas	3	3	2	2	4	5	4	5	5	6	8	10	10
Holding Company - Petrochemicals and Rubber	3	3	2	2	1	1	1	1	1	1	1	1	1
Holding Company - Pharmaceuticals and Hygiene	1	1	1	0	0	0	1	1	1	1	0	0	0
Holding Company - Pulp and Paper	1	1	1	1	1	1	1	1	1	1	1	1	1
Holding Company - Real Estate Credit	4	4	4	4	4	4	4	3	3	3	3	3	3
Holding Company - Sanitization and Utilities	2	2	2	3	2	2	2	2	3	4	4	5	5
Holding Company - Securities	4	4	2	2	2	2	2	2	2	2	2	2	2
Holding Company - Stock Exchange	1	0	0	0	0	0	0	0	0	0	0	0	0
Holding Company - Telecommunications	18	18	17	18	14	12	11	9	5	5	6	6	6
Holding Company - Textile Industries	3	3	2	2	2	2	2	2	2	2	3	2	2
Holding Company - Toys and Leisure	1	1	1	1	1	1	1	1	1	1	2	2	1
Holding Company - Transport and Logistics	15	16	16	14	12	12	13	13	14	13	14	13	13
Total	**202**	**213**	**195**	**184**	**178**	**170**	**166**	**168**	**154**	**152**	**176**	**176**	**169**

Table 11 – Summary - Average length (number of pages) of the financial reports per Sector per Year

Sector	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Agriculture	63	78	78	82	78	77	77	81	83	84	86	98	100
Banking	75	79	86	89	89	91	88	89	87	83	89	114	113
Civil Construction	61	69	73	71	70	69	65	68	67	67	72	75	78
Commerce	83	83	83	76	71	71	69	77	79	78	80	83	85
Communication and Informatics	70	65	67	68	72	73	70	75	80	73	82	81	85
Drinks and Tabacco	78	94	109	105	99	117	120	125	136	138	135	120	108
Education	31	31	65	64	62	60	65	66	75	85	84	76	71
Electricity	69	72	75	75	75	76	78	78	81	80	78	74	72
Factoring	29	0	0	0	0	0	0	0	0	0	0	0	0
Financial Intermediary	71	67	71	64	67	64	57	55	54	56	66	78	85
Food	66	78	75	77	77	83	72	71	71	73	76	82	84
Graphical Design and Publishing	81	83	85	83	80	70	68	81	77	0	0	0	0
Hospitality and Tourism	30	35	41	42	40	38	38	41	42	44	40	38	40
Insurance	104	90	84	82	70	66	85	84	83	92	97	100	103
Leasing	32	36	36	37	37	36	35	35	36	37	37	36	33
Machines, Equipment, Vehicles and Parts	63	65	68	65	63	61	65	67	70	71	71	73	72
Medical Services	86	95	95	94	90	88	88	87	86	87	83	84	85
Metallurgy and Steel	48	53	52	53	53	52	52	53	57	59	60	60	62
Mineral Extraction	85	78	77	75	61	56	54	57	59	64	73	73	72
Oil and Gas	39	34	39	46	44	43	40	39	45	52	54	60	60
Packaging	84	72	93	97	95	72	88	88	67	71	98	98	101
Petrochemicals and Rubber	71	70	68	72	69	71	65	68	65	71	76	73	75
Pharmaceuticals and Hygiene	80	69	67	74	80	76	79	76	67	75	78	72	74
Pulp and Paper	70	68	75	76	73	74	79	83	85	91	101	106	96
Real Estate Credit	27	35	49	29	28	30	30	30	31	29	31	33	34
Reforestation	23	23	22	22	19	19	19	22	21	20	20	20	21
Sanitization and Utilities	65	63	66	69	66	67	65	67	68	73	79	79	84
Securities	26	29	29	29	29	30	30	31	30	30	29	30	33
Stock Exchange	111	101	94	79	82	81	99	89	88	80	83	87	88
Telecommuncations	82	76	86	81	75	79	77	83	89	95	92	90	89
Textile Industries	61	61	63	63	62	62	62	63	66	70	74	76	79
Toys and Leisure	45	46	46	45	42	43	49	57	60	60	69	69	67
Transport and Logistics	60	59	60	59	58	54	53	56	57	60	60	62	60
Total	**63**	**62**	**66**	**65**	**63**	**62**	**63**	**65**	**65**	**65**	**68**	**70**	**70**

Table 12 – Summary - Average length (number of pages) of the financial reports per Sector per Year (Holding Companies)

Sector	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Holding Company - Agriculture	60	78	97	94	77	81	81	87	96	88	81	81	73
Holding Company - Civil Construction	84	88	99	84	84	83	72	81	85	81	85	86	90
Holding Company - Commerce	69	70	75	65	69	63	62	71	77	82	80	84	80
Holding Company - Communication and Informatics	0	0	29	38	55	59	63	57	67	73	56	63	59
Holding Company - Education	86	66	63	73	59	62	62	65	68	70	76	77	76
Holding Company - Electricity	74	72	75	76	78	80	79	72	76	71	73	69	67
Holding Company - Financial Intermediary	53	50	51	49	47	47	47	46	43	42	46	59	68
Holding Company - Food	73	74	76	72	72	77	75	77	75	73	71	67	70
Holding Company - Graphical Design and Publishing	50	36	32	33	33	33	0	0	0	0	0	0	0
Holding Company - Hospitality and Tourism	56	63	80	76	57	51	25	31	44	50	50	46	49
Holding Company - Insurance	81	85	112	108	90	85	81	85	90	93	101	98	91
Holding Company - Leasing	35	33	22	20	20	20	0	0	0	0	0	0	0
Holding Company - Machines, Equipment, Vehicles and Parts	66	65	67	62	56	56	61	65	71	76	75	75	73
Holding Company - Medical Services	40	21	54	84	88	0	0	90	101	111	101	94	93
Holding Company - Metallurgy and Steel	70	74	72	75	65	64	65	69	66	66	71	73	69
Holding Company - Mineral Extraction	52	43	46	51	48	50	53	52	47	46	42	44	44
Holding Company - No Main Sector	38	41	40	41	38	39	40	39	40	44	45	47	47
Holding Company - Oil and Gas	62	75	96	99	77	73	85	78	85	95	103	95	92
Holding Company - Petrochemicals and Rubber	75	81	83	81	88	100	110	123	112	118	96	103	106
Holding Company - Pharmaceuticals and Hygiene	65	28	27	0	0	0	43	64	57	51	0	0	0
Holding Company - Pulp and Paper	74	83	75	65	69	72	77	81	85	89	89	87	86
Holding Company - Real Estate Credit	40	41	38	35	31	31	31	35	32	35	34	37	42
Holding Company - Sanitization and Utilities	59	64	64	57	67	71	76	83	84	89	84	78	87
Holding Company - Securities	19	23	30	29	34	33	33	32	29	30	30	33	36
Holding Company - Stock Exchange	54	0	0	0	0	0	0	0	0	0	0	0	0
Holding Company - Telecommunications	65	57	53	57	49	43	41	45	56	59	66	70	72
Holding Company - Textile Industries	58	67	58	62	63	62	61	67	73	71	63	74	84
Holding Company - Toys and Leisure	36	37	39	48	48	48	54	48	48	47	50	50	43
Holding Company - Transport and Logistics	65	62	60	61	63	60	68	69	73	85	86	83	83
Total	**57**	**54**	**59**	**58**	**56**	**53**	**53**	**59**	**61**	**63**	**60**	**61**	**61**

Table 13 – Summary - Average size (in megabytes) of the financial reports per Sector per Year

Sector	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Agriculture	4,68	5,98	7,47	8,96	11,3	9,36	9,21	9,91	7,9	8,89	8,78	8,31	9,28
Banking	4,6	4,61	5,26	5,59	5,29	5,68	5,16	6,31	6,13	6,3	7,05	8,96	9,92
Civil Construction	5,23	5,99	6,57	6,54	6,71	6,64	6,84	7,01	7,6	7,58	8,11	8,25	10,4
Commerce	6,06	6,95	6,94	7,07	7,09	6,7	7,1	7,12	9,28	8,31	8,68	8,62	8,78
Communication and Informatics	5,99	6,11	6,69	6,95	6,87	6,87	6,39	7,12	7,52	7,02	7,71	7,98	8,54
Drinks and Tabacco	5,13	6	8,3	8,01	7,39	8	8	8	8,01	10,7	8	8	6,67
Education	4,08	4,09	5,39	5,7	5,75	5,54	7,04	6,62	8,42	8,29	9,03	8,57	7,11
Electricity	4,49	4,78	4,96	5,13	5,21	5,4	5,89	5,59	6,09	7,07	6,33	6,61	6,07
Factoring	2,64	0	0	0	0	0	0	0	0	0	0	0	0
Financial Intermediary	6,48	6,18	6,18	6,73	6,18	6,18	5,66	5,2	7,45	5,27	6,51	7,28	7,02
Food	5,83	6,77	6,09	6,38	6,96	7,67	6,64	7,06	7,05	7,18	8,11	8,09	7,82
Graphical Design and Publishing	4	4	5,33	8	8	8	7,39	8	8	0	0	0	0
Hospitality and Tourism	3,12	3,08	3,81	3,62	3,67	4,47	4,05	4,5	4,94	5,74	5,73	5,66	5,4
Insurance	6,39	7,27	7,23	7,27	7,11	6,96	7,35	7,27	8,72	16,4	8	8	7,88
Leasing	3,09	3,02	2,92	3,21	3,61	3,75	3,71	3,63	3,7	3,73	3,45	3,32	3,55
Machines, Equipment, Vehicles and Parts	5,94	5,85	5,82	5,77	5,66	5,66	6,29	6,29	7,24	6,8	6,66	7,23	6,68
Medical Services	5	7,67	8,4	8	8	9,6	8	9,14	7,67	9	8,56	10,4	9,03
Metallurgy and Steel	4,35	5,41	5,94	5,44	5,37	5,67	5,81	6,19	6,72	6,48	7,49	6,6	7,23
Mineral Extraction	6,21	6,2	7,09	6,2	5,75	6,2	5,84	6,45	6,43	6,92	7,2	9,2	6,79
Oil and Gas	3,82	3,61	4,2	4,39	4,77	4,78	5,15	5,14	6,02	6,23	6,38	6,42	7,61
Packaging	5,14	5,75	6,92	6,62	6,93	5,79	6,29	6,99	5,54	6,22	11,3	14	12
Petrochemicals and Rubber	6,34	6,46	6,03	6,16	6,27	6,41	6,47	6,75	6,41	7,09	6,85	7,83	7,78
Pharmaceuticals and Hygiene	5,77	7,43	5,98	6,51	7,78	7,01	7,12	6,97	6,27	6,89	7,78	8,65	7,02
Pulp and Paper	5,37	5,93	6,4	6,25	5,79	7,24	7,38	6,71	7,68	7,17	8,45	8,5	9,56
Real Estate Credit	2,61	2,71	5,81	2,65	2,63	2,66	2,66	2,67	2,67	2,64	2,67	2,7	2,71
Reforestation	2,55	2,57	2,56	2,55	2,52	2,51	2,51	2,56	2,54	4,22	2,53	2,53	2,54
Sanitization and Utilities	4,05	4,11	4,3	4,44	4,34	4,46	4,6	4,87	5,27	5,48	6,09	6,88	10,9
Securities	2,82	2,99	3,14	3,55	3,44	3,79	3,96	4,1	4,26	4,64	4,43	4,25	4,15
Stock Exchange	8	8	8	8	8	8	8	8	8	8	8	8	8
Telecommuncations	5,43	5,51	6,21	6,01	6,05	6,29	7,09	6,83	6,28	6,81	6,88	7,06	7,5
Textile Industries	4,97	5,27	5,8	5,98	6,16	6,85	6,25	6,75	8,96	8,73	9,35	9,88	10,9
Toys and Leisure	4,51	4,54	4,14	4,31	4,28	6,78	6,51	7,11	7,65	6,33	7,21	6,99	7,29
Transport and Logistics	4,19	5,06	4,57	4,32	4,6	4,46	4,58	4,92	5,77	5,52	5,8	5,78	5,92
Total	**4,81**	**5,15**	**5,59**	**5,65**	**5,74**	**5,92**	**5,91**	**6,11**	**6,43**	**6,6**	**6,64**	**6,99**	**7,03**

Table 14 – Summary - Average size (in megabytes) of the financial reports per Sector per Year (Holding Companies)

Sector	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Holding Company - Agriculture	6,69	10,7	8	8	8	8	8	8	8	8	12	16	8
Holding Company - Civil Construction	5,57	6,29	6,6	6,77	7,2	7,34	6,38	7,21	10,7	9,08	11	8,54	9,57
Holding Company - Commerce	5,46	5,93	5,84	5,71	5,84	5,96	6,88	6,64	8,08	7,06	10,8	11,3	9,82
Holding Company - Communication and Informatics	0	0	5,26	4,63	5,95	6,05	6,15	5,98	7,37	8,42	8,4	6,77	7,53
Holding Company - Education	5,63	4,85	5	6,25	6,58	6,82	6,83	6,83	7,02	6,73	7,01	7,12	7,1
Holding Company - Electricity	5,31	5,22	5,27	5,66	5,77	6,06	5,9	6,05	7,13	7,17	7,35	8,01	6,71
Holding Company - Financial Intermediary	4,88	5,23	5,17	5,05	5,58	5,39	5,48	6,36	5,42	5,62	6,72	6,63	8,05
Holding Company - Food	5,89	7,59	7,17	6,38	6,51	6,64	6,44	6,49	6,52	6,58	6,6	6,54	7,45
Holding Company - Graphical Design and Publishing	5,76	5,44	5,34	2,67	8,87	2,67	0	0	0	0	0	0	0
Holding Company - Hospitality and Tourism	9,53	7,79	7,73	8,68	6,36	5,81	2,59	3,57	4,31	5,85	12,7	5,69	5,76
Holding Company - Insurance	5,8	7,56	8,02	8	7,82	7,82	7,83	7,84	11,2	10	9,33	12	8,89
Holding Company - Leasing	4,17	3,87	2,54	2,52	2,53	2,53	0	0	0	0	0	0	0
Holding Company - Machines, Equipment, Vehicles and Parts	6,37	6,36	7,15	6,77	6,27	5,98	7,74	7,1	7,76	7,25	8,94	9,92	10,8
Holding Company - Medical Services	5,55	5,08	4,37	8	10,7	0	0	24	8	9,33	8	8	8
Holding Company - Metallurgy and Steel	5,9	5,93	5,67	5,8	5,37	5,26	5,3	5,74	5,79	7,37	7,24	6,84	7,62
Holding Company - Mineral Extraction	5,53	4,58	5,23	5,6	5,54	5,98	5,94	5,6	5,08	5,29	4,96	5,02	5,64
Holding Company - No Main Sector	3,72	3,9	3,89	4	3,79	4,03	4,43	4,44	4,57	5,08	4,82	5,08	5,71
Holding Company - Oil and Gas	5,96	6,36	8	8	7,08	6,69	7,06	6,77	7,16	6,87	9,24	8,02	9,11
Holding Company - Petrochemicals and Rubber	6,59	7,71	8	6,4	6,67	8	8	8	8	8	8	34,7	8
Holding Company - Pharmaceuticals and Hygiene	4,08	2,63	2,62	0	0	0	2,81	4,59	9,94	2,91	0	0	0
Holding Company - Pulp and Paper	4	4	4	5,43	4,71	4	6,67	6,67	8	8	8	8	8
Holding Company - Real Estate Credit	5,58	5,22	4,92	4,01	4,09	4,27	4,07	3,7	3,65	3,7	3,99	5,32	5,64
Holding Company - Sanitization and Utilities	4,17	3,48	4,42	4,81	4,47	6,95	7,01	6,99	7,3	6,25	6,88	6,35	6,19
Holding Company - Securities	2,51	2,57	2,65	2,64	3,13	2,7	2,69	6,13	3,07	3,09	3,12	4,1	4,83
Holding Company - Stock Exchange	5,93	0	0	0	0	0	0	0	0	0	0	0	0
Holding Company - Telecommunications	5,58	5,09	4,69	4,78	4,15	4,52	4,37	4,12	5,39	5,76	6,03	6,55	7,05
Holding Company - Textile Industries	5,25	5,07	6,8	6,86	6,87	7,79	6,85	6,88	16,7	18	12,7	6,87	17,3
Holding Company - Toys and Leisure	2,71	2,72	2,76	2,87	3,82	2,87	2,93	5,74	7,64	5,73	5,81	7,34	5,63
Holding Company - Transport and Logistics	4,7	5,03	5,17	5,56	5,63	5,58	5,9	6	6,3	6,78	7,81	7,58	7,58
Total	**5,13**	**5,04**	**5,25**	**5,24**	**5,49**	**5,02**	**4,97**	**6,12**	**6,56**	**6,34**	**6,81**	**7,52**	**6,76**

Table 15 – Average score, per year, per sector - Informativeness Index

Sector	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Agriculture	50,04	53,18	51,63	49,58	50,28	50,68	49,26	49,32	49,19	52,01	52,64	52,07	52,09
Banking	47,15	48,68	49,27	49,42	49,37	49,67	49,42	50,16	50,07	51,54	51,01	53,25	52,33
Civil Construction	44,41	46,48	46,11	45,36	45,70	46,07	46,79	46,93	46,99	48,56	47,67	47,34	47,47
Commerce	47,06	48,21	47,40	48,01	47,48	48,09	47,89	48,26	49,34	52,23	51,57	51,39	51,36
Communication and Informatics	44,93	45,08	47,03	45,56	48,27	46,31	46,86	47,31	49,11	52,00	52,07	51,84	50,81
Drinks and Tabacco	49,83	51,86	56,25	52,62	50,56	50,85	51,67	55,19	55,54	55,77	52,32	51,78	51,96
Education	38,87	39,40	45,40	48,64	49,81	47,97	48,18	49,71	50,74	52,52	49,53	50,30	50,15
Electricity	48,34	49,86	51,06	50,59	50,14	51,06	51,49	51,17	50,76	52,16	51,63	51,30	50,64
Factoring	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Financial Intermediary	46,67	45,41	45,09	46,59	45,76	46,41	44,48	44,35	43,83	45,64	44,63	46,32	42,20
Food	46,86	49,99	48,66	49,18	50,93	50,64	49,78	49,78	50,58	52,13	52,50	51,17	51,57
Graphical Design and Publishing	57,78	53,53	54,26	52,18	53,19	52,89	47,22	52,43	55,77	0,00	0,00	0,00	0,00
Hospitality and Tourism	38,25	42,41	42,57	42,15	45,31	44,10	44,22	45,58	45,77	46,76	47,31	42,78	45,66
Insurance	48,47	46,94	43,65	43,98	44,91	43,52	46,99	46,88	48,31	50,31	50,81	50,20	50,77
Leasing	46,33	46,97	46,81	47,54	46,89	46,86	48,48	48,03	48,77	48,27	44,52	43,56	41,53
Machines, Equipment, Vehicles and Parts	45,49	46,93	47,38	47,59	47,40	47,44	47,84	47,98	48,81	50,00	48,88	49,47	48,69
Medical Services	45,54	47,71	48,18	48,67	47,34	49,21	50,74	50,30	49,07	51,54	52,42	52,49	53,89
Metallurgy and Steel	44,46	47,37	47,86	47,61	47,69	47,91	48,17	48,46	48,59	48,83	48,37	48,64	48,16
Mineral Extraction	42,19	44,05	45,36	45,72	45,25	44,25	44,01	41,43	40,85	47,85	52,68	50,53	49,81
Oil and Gas	42,70	41,18	40,94	35,80	34,19	37,98	35,63	37,72	39,34	43,23	43,90	45,06	46,77
Packaging	46,99	46,88	45,10	45,76	46,30	48,42	48,84	48,42	49,25	50,77	51,90	49,77	44,34
Petrochemicals and Rubber	50,12	48,24	46,04	46,91	49,13	49,38	48,09	48,19	48,77	51,25	48,96	49,56	49,82
Pharmaceuticals and Hygiene	45,57	47,98	50,44	48,40	47,77	47,29	48,81	48,09	47,75	51,34	49,72	49,97	50,64
Pulp and Paper	49,73	50,98	50,41	50,48	49,16	49,85	49,35	48,70	48,79	50,68	52,80	53,11	54,39
Real Estate Credit	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Reforestation	46,77	46,20	44,57	43,80	46,36	48,14	47,94	50,68	52,73	48,61	46,97	46,43	44,16
Sanitization and Utilities	46,26	47,88	48,40	49,07	48,42	48,14	48,02	49,10	50,18	53,78	52,06	50,16	49,31
Securities	37,95	38,95	38,81	38,24	37,70	38,19	38,47	39,16	38,76	39,80	38,99	39,66	40,16
Stock Exchange	51,72	53,98	52,71	55,85	54,67	56,28	56,46	51,35	53,16	54,09	54,05	52,09	53,44
Telecommuncations	44,10	45,21	45,60	44,89	44,70	44,53	46,50	46,97	47,50	48,82	48,76	48,61	48,92
Textile Industries	43,59	44,08	45,01	45,32	45,73	45,81	44,96	45,70	47,33	48,29	48,26	47,41	46,75
Toys and Leisure	42,43	42,98	42,76	42,91	42,98	41,58	43,00	46,45	45,22	45,43	48,46	47,01	47,01
Transport and Logistics	45,89	47,61	46,27	46,27	47,64	47,34	47,07	48,16	48,77	50,76	49,96	49,40	48,80
Total	**43,23**	**44,13**	**44,27**	**44,08**	**44,27**	**44,45**	**44,44**	**44,91**	**45,44**	**45,30**	**45,01**	**44,63**	**44,35**

Table 16 – Average score, per year, per sector - Informativeness Index (Holdings)

Sector	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Holding Company - Agriculture	51,19	53,75	59,77	60,55	47,92	48,23	49,67	50,81	50,98	55,06	54,41	54,33	50,27
Holding Company - Civil Construction	42,01	43,76	45,21	44,52	43,04	43,34	44,45	47,28	47,43	46,36	47,67	47,63	47,64
Holding Company - Commerce	43,45	47,57	44,56	45,60	43,75	43,41	44,08	44,97	45,82	50,68	47,40	48,91	49,99
Holding Company - Communication and Informatics	0,00	0,00	39,51	41,59	55,03	47,56	57,52	53,02	51,62	49,81	45,19	47,25	47,52
Holding Company - Education	46,38	46,81	44,91	48,81	47,90	47,08	46,35	46,61	46,08	47,65	45,94	47,97	47,77
Holding Company - Electricity	43,70	44,52	45,37	44,61	44,71	45,67	45,07	46,85	45,93	46,42	47,91	47,98	46,90
Holding Company - Financial Intermediary	41,56	43,64	45,78	43,98	42,65	41,67	43,28	45,75	42,92	44,00	43,40	43,00	42,19
Holding Company - Food	48,62	48,06	47,86	48,41	49,86	47,87	47,01	49,17	50,58	50,95	50,96	51,08	52,27
Holding Company - Graphical Design and Publishing	43,41	51,41	44,11	48,78	49,68	41,59	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Holding Company - Hospitality and Tourism	43,99	43,75	46,40	46,32	43,37	42,74	42,93	44,17	46,30	48,53	42,36	43,11	42,74
Holding Company - Insurance	51,06	49,65	47,67	46,82	46,44	49,21	49,63	50,08	51,89	53,36	52,33	51,68	49,91
Holding Company - Leasing	41,66	41,20	31,79	29,81	30,12	31,38	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Holding Company - Machines, Equipment, Vehicles and Parts	44,26	46,74	47,44	50,43	54,14	51,71	49,96	51,26	50,56	54,76	50,68	49,69	52,61
Holding Company - Medical Services	60,33	61,68	57,28	56,48	52,21	55,32	0,00	55,03	52,92	54,55	52,92	54,04	59,77
Holding Company - Metallurgy and Steel	45,40	47,64	48,53	48,96	48,00	47,60	47,08	51,54	52,42	49,98	50,04	50,23	49,38
Holding Company - Mineral Extraction	42,26	42,37	40,91	40,32	40,87	42,24	44,12	44,66	49,52	51,09	47,63	46,43	46,04
Holding Company - No Main Sector	39,56	41,82	39,52	39,69	39,06	39,91	41,20	40,87	40,74	42,60	42,72	43,57	42,72
Holding Company - Oil and Gas	39,89	42,50	44,77	42,75	45,47	46,44	47,67	48,72	47,27	50,86	53,03	52,40	51,35
Holding Company - Petrochemicals and Rubber	47,58	50,75	53,34	56,02	57,16	57,88	57,69	58,49	53,59	58,68	56,59	53,55	55,59
Holding Company - Pharmaceuticals and Hygiene	0,00	36,51	36,00	0,00	0,00	0,00	53,43	55,99	57,66	61,26	0,00	0,00	0,00
Holding Company - Pulp and Paper	49,98	56,24	51,02	49,56	48,70	48,77	49,52	50,72	58,13	60,71	55,29	56,01	56,49
Holding Company - Real Estate Credit	40,94	41,01	40,70	38,74	37,08	39,30	42,12	45,47	42,89	42,68	42,23	42,19	40,44
Holding Company - Sanitization and Utilities	47,62	48,60	49,86	46,87	46,65	46,88	48,03	49,92	52,38	53,46	53,46	52,92	50,85
Holding Company - Securities	37,67	43,85	44,37	40,51	43,04	43,28	41,71	44,70	36,16	41,92	44,26	46,64	47,04
Holding Company - Stock Exchange	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Holding Company - Telecommunications	42,66	42,76	43,22	45,32	43,67	41,30	41,08	42,02	44,92	47,42	45,66	49,27	48,28
Holding Company - Textile Industries	43,22	49,67	42,66	41,32	41,94	47,54	42,44	44,97	45,02	44,89	44,08	42,54	53,75
Holding Company - Toys and Leisure	39,56	41,56	46,26	40,41	41,30	39,46	41,48	48,54	50,24	47,14	45,09	49,19	38,70
Holding Company - Transport and Logistics	45,45	45,75	44,03	44,02	45,41	43,75	45,42	45,30	46,83	51,16	51,27	50,11	49,00
Total	**40,12**	**43,23**	**43,89**	**42,45**	**42,39**	**42,11**	**40,10**	**43,34**	**43,48**	**45,03**	**41,81**	**42,13**	**42,04**

Table 17 – Average score, per year, per sector - Flesch Reading Ease

Sector	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Agriculture	66,54	72,52	81,50	85,70	90,93	87,09	85,81	87,40	84,32	84,91	86,14	82,23	82,11
Banking	78,82	77,93	77,66	78,07	78,72	80,36	79,57	78,87	78,39	78,32	77,39	75,79	74,68
Civil Construction	76,90	75,32	75,50	77,54	79,43	79,96	79,64	79,20	80,05	75,95	77,17	79,81	80,41
Commerce	74,85	75,90	78,35	80,88	82,12	82,79	82,50	80,37	82,28	78,03	79,14	80,03	80,71
Communication and Informatics	70,58	72,00	79,70	76,94	81,59	81,17	82,00	87,54	85,01	79,89	79,99	78,69	78,23
Drinks and Tabacco	79,83	80,54	70,69	92,63	93,80	97,54	97,40	90,16	91,24	92,78	92,91	84,31	81,89
Education	64,84	62,97	67,35	70,58	71,39	72,68	75,42	77,87	78,97	74,88	77,36	79,86	83,96
Electricity	74,98	74,93	76,03	75,26	76,41	79,34	78,85	77,29	79,22	75,51	75,37	77,50	77,41
Factoring	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Financial Intermediary	68,38	67,50	68,26	69,28	69,76	71,24	67,01	68,07	63,56	59,97	63,78	64,62	62,57
Food	71,47	68,60	70,56	65,65	70,52	73,58	82,41	82,74	83,08	80,98	79,67	84,48	83,32
Graphical Design and Publishing	76,99	84,52	82,62	83,16	84,29	80,90	84,77	83,25	83,50	0,00	0,00	0,00	0,00
Hospitality and Tourism	66,39	68,17	67,49	65,39	76,77	81,60	82,05	85,30	86,10	81,46	80,76	82,67	80,10
Insurance	71,24	74,94	68,28	70,43	74,62	77,69	85,27	85,49	87,77	84,29	87,38	81,65	83,50
Leasing	70,52	72,21	71,25	68,93	67,95	66,60	67,23	67,31	69,97	69,63	59,16	63,28	68,23
Machines, Equipment, Vehicles and Parts	76,53	77,42	76,74	77,12	78,09	78,65	79,31	77,79	78,87	78,56	78,85	79,92	81,89
Medical Services	78,21	78,41	78,41	82,22	82,93	83,94	81,95	83,66	83,16	81,60	81,56	81,12	83,51
Metallurgy and Steel	76,44	76,01	76,62	77,27	78,91	79,48	79,29	78,62	82,03	78,60	80,91	78,82	78,83
Mineral Extraction	68,44	70,15	72,51	71,92	71,71	73,27	71,92	70,91	61,59	65,24	69,04	70,25	78,10
Oil and Gas	59,03	61,73	65,14	69,93	70,78	68,92	69,17	70,18	68,00	66,16	66,02	68,84	74,78
Packaging	75,78	74,75	73,78	73,73	73,53	72,13	68,26	66,26	67,69	68,02	66,37	73,71	78,05
Petrochemicals and Rubber	81,22	82,49	82,81	81,81	81,24	81,04	82,49	82,33	82,55	79,41	82,73	81,95	81,37
Pharmaceuticals and Hygiene	68,66	63,65	65,94	69,88	71,92	72,49	71,97	76,32	74,97	80,30	79,25	78,78	79,22
Pulp and Paper	74,53	71,66	75,41	78,87	80,27	79,07	81,94	80,48	77,85	75,67	80,81	82,94	90,07
Real Estate Credit	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Reforestation	70,81	63,59	67,40	66,17	65,62	64,49	62,69	59,62	65,01	65,22	65,55	66,82	75,34
Sanitization and Utilities	72,59	72,62	72,07	70,72	73,16	72,71	73,04	74,23	74,47	73,32	74,60	75,91	76,66
Securities	58,52	57,57	57,88	57,24	58,44	60,85	62,89	61,34	55,56	55,20	55,75	55,41	59,53
Stock Exchange	75,24	75,07	74,06	83,35	81,25	79,38	82,24	84,49	89,05	84,83	90,34	85,89	100,00
Telecommuncations	75,80	72,79	73,36	72,68	73,34	75,11	74,00	74,28	75,10	71,82	75,91	76,18	78,21
Textile Industries	76,80	77,42	76,84	77,68	79,84	79,25	77,38	77,17	78,05	74,90	77,65	78,71	79,34
Toys and Leisure	72,48	69,00	69,06	66,55	66,85	64,37	73,71	73,31	71,68	70,96	72,78	76,68	75,29
Transport and Logistics	70,76	73,64	73,84	74,18	75,71	75,83	76,10	76,11	77,47	75,59	77,09	78,06	78,55
Total	**68,01**	**68,06**	**68,70**	**70,05**	**71,57**	**71,92**	**72,67**	**72,66**	**72,62**	**68,55**	**69,44**	**69,85**	**71,69**

Table 18 – Average score, per year, per sector - Flesch Reading Ease (Holdings)

Sector	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Holding Company - Agriculture	73,46	75,48	67,79	80,24	79,75	82,79	83,50	78,95	79,92	73,89	72,37	73,36	64,65
Holding Company - Civil Construction	74,56	78,40	79,69	81,06	81,25	82,52	80,33	76,26	81,44	75,76	75,92	80,50	82,03
Holding Company - Commerce	60,98	70,73	66,38	71,13	75,56	75,70	76,68	73,87	75,81	69,75	72,25	72,18	77,21
Holding Company - Communication and Informatics	0,00	0,00	100,00	89,73	76,15	83,94	82,09	80,70	80,18	78,26	91,02	74,73	74,27
Holding Company - Education	73,35	72,61	71,78	77,05	69,83	72,55	71,55	69,45	70,15	74,78	71,98	77,38	78,99
Holding Company - Electricity	69,94	66,08	68,65	70,56	71,91	75,29	74,16	72,63	76,64	71,40	73,10	77,05	78,36
Holding Company - Financial Intermediary	69,33	71,72	72,54	71,01	71,09	63,95	79,05	78,61	73,13	80,50	72,61	67,67	69,43
Holding Company - Food	78,91	80,11	85,41	89,33	91,90	90,17	88,77	89,32	88,32	86,89	87,08	90,73	87,24
Holding Company - Graphical Design and Publishing	60,13	68,68	69,54	47,63	48,10	47,12	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Holding Company - Hospitality and Tourism	79,74	75,63	78,70	88,86	96,37	100,00	66,65	65,96	67,98	66,71	67,47	66,72	66,93
Holding Company - Insurance	80,88	83,63	80,07	84,54	77,66	90,29	93,04	89,32	90,97	79,18	76,01	80,69	78,91
Holding Company - Leasing	71,07	73,95	71,16	73,12	71,00	67,64	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Holding Company - Machines, Equipment, Vehicles and Parts	69,87	73,41	72,41	72,62	75,06	76,74	77,75	76,39	79,95	74,45	76,32	81,01	81,80
Holding Company - Medical Services	59,80	67,85	59,58	62,74	61,60	54,80	0,00	70,03	79,79	73,36	76,65	80,52	71,88
Holding Company - Metallurgy and Steel	76,43	77,16	76,04	75,55	77,27	81,76	81,75	78,77	80,50	77,50	78,11	78,72	82,87
Holding Company - Mineral Extraction	70,70	67,45	70,11	72,37	72,73	75,50	75,87	73,05	63,20	61,52	61,63	66,20	68,63
Holding Company - No Main Sector	63,15	62,38	65,22	66,11	65,87	66,27	66,60	66,68	67,26	66,06	67,28	67,43	67,49
Holding Company - Oil and Gas	66,08	66,02	63,05	61,87	64,93	73,07	75,90	76,27	76,16	76,09	78,91	79,26	78,17
Holding Company - Petrochemicals and Rubber	76,99	79,75	81,27	77,94	82,49	83,80	82,52	77,35	85,67	79,30	82,15	83,44	81,04
Holding Company - Pharmaceuticals and Hygiene	0,00	48,10	54,65	0,00	0,00	0,00	74,45	75,87	71,80	67,43	0,00	0,00	0,00
Holding Company - Pulp and Paper	81,40	84,91	89,95	96,66	97,08	95,42	90,59	87,87	86,20	82,53	81,53	80,50	82,06
Holding Company - Real Estate Credit	62,08	57,85	60,37	59,65	63,02	66,66	65,09	58,75	62,99	57,19	61,51	69,48	64,81
Holding Company - Sanitization and Utilities	73,17	73,50	75,40	72,90	84,00	84,53	87,43	84,10	90,66	85,12	88,81	88,84	88,27
Holding Company - Securities	58,42	54,88	53,75	57,16	61,23	61,46	63,06	60,96	58,67	58,77	51,58	49,22	51,23
Holding Company - Stock Exchange	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Holding Company - Telecommunications	66,61	68,35	67,54	66,47	67,10	67,75	65,00	64,27	70,89	71,76	73,69	76,65	71,92
Holding Company - Textile Industries	76,62	74,08	72,79	73,95	78,53	77,99	75,72	75,18	80,09	81,79	75,40	77,33	76,74
Holding Company - Toys and Leisure	56,61	57,05	59,11	73,19	69,61	73,46	75,73	75,82	74,83	51,48	59,32	78,98	54,17
Holding Company - Transport and Logistics	71,78	71,07	71,72	75,47	76,73	75,93	76,13	73,55	75,58	75,28	75,22	77,72	79,34
Total	**62,83**	**65,55**	**69,13**	**68,58**	**69,23**	**70,59**	**66,53**	**67,24**	**68,58**	**65,41**	**63,72**	**65,39**	**64,08**

Table 19 – Average score, per year, per sector - Gunning-Fog Index

Sector	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**	**2023**
Agriculture	17,40	17,62	15,37	13,98	13,06	13,81	14,08	13,82	14,41	14,15	14,09	14,65	14,69
Banking	14,91	15,14	15,22	15,21	15,35	14,75	14,74	15,15	19,68	15,38	15,49	15,67	17,64
Civil Construction	15,92	16,11	15,98	15,66	15,28	15,35	15,81	15,85	15,53	16,60	15,86	15,27	14,91
Commerce	16,19	15,77	15,44	14,95	14,65	14,54	14,67	15,04	14,81	16,12	15,16	14,90	14,85
Communication and Informatics	17,24	18,68	17,09	18,20	14,89	14,96	14,80	13,71	14,98	15,03	15,14	15,23	15,08
Drinks and Tabacco	15,98	15,69	17,48	12,84	12,41	12,00	12,00	13,25	12,97	12,84	12,68	14,18	14,58
Education	18,55	19,12	17,85	17,43	17,79	17,42	16,52	16,04	15,49	16,29	15,81	15,32	14,64
Electricity	15,71	15,71	15,60	15,38	15,24	14,97	15,23	15,77	15,19	16,02	16,03	15,69	15,56
Factoring	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Financial Intermediary	17,53	17,81	17,91	17,66	17,72	17,49	18,38	18,25	19,13	19,82	18,67	18,35	18,90
Food	16,15	16,37	16,16	16,15	15,83	15,70	15,07	15,04	15,01	15,25	15,25	14,33	14,39
Graphical Design and Publishing	15,90	14,60	15,00	14,76	14,59	14,89	13,96	14,35	14,24	0,00	0,00	0,00	0,00
Hospitality and Tourism	18,04	18,11	17,78	22,99	15,81	14,80	14,69	14,22	14,11	14,76	14,94	14,09	15,03
Insurance	16,85	16,19	17,96	17,17	16,14	15,39	14,34	14,13	14,21	14,53	14,41	15,10	14,70
Leasing	16,59	16,59	16,78	17,29	17,58	18,76	17,34	17,23	16,87	17,39	21,01	18,66	17,55
Machines, Equipment, Vehicles and Parts	15,73	15,64	15,72	15,70	15,69	15,28	15,35	15,63	15,44	15,61	15,46	15,27	14,89
Medical Services	15,39	15,28	15,37	14,91	14,73	14,59	14,91	14,71	14,43	14,49	14,73	16,67	14,66
Metallurgy and Steel	16,38	15,97	15,81	15,64	15,42	15,23	15,22	15,39	14,80	15,42	15,07	15,45	15,39
Mineral Extraction	17,83	17,40	16,82	16,98	17,01	16,66	16,96	17,21	19,42	18,90	17,88	20,23	16,10
Oil and Gas	19,74	19,50	18,81	17,66	17,41	17,74	17,60	17,91	18,34	19,17	18,79	18,13	16,58
Packaging	15,83	16,01	16,29	16,25	16,17	16,78	17,53	18,00	17,68	17,56	18,05	16,69	15,84
Petrochemicals and Rubber	14,90	14,69	14,62	14,75	14,87	14,99	14,81	14,80	14,64	15,47	14,96	15,06	14,95
Pharmaceuticals and Hygiene	17,53	20,11	19,49	17,28	16,94	16,71	18,37	16,09	16,68	15,37	15,43	15,50	15,36
Pulp and Paper	16,31	16,87	15,99	15,21	14,99	15,30	14,61	14,92	15,66	15,94	15,21	14,50	14,14
Real Estate Credit	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Reforestation	17,22	18,93	17,93	18,47	18,46	18,89	19,24	19,79	18,72	18,47	18,37	17,99	16,99
Sanitization and Utilities	16,90	16,89	16,98	17,31	16,69	16,89	16,90	16,77	16,66	17,17	16,62	16,30	16,17
Securities	19,52	19,72	19,88	20,08	19,72	19,29	19,04	20,18	23,07	20,65	20,73	21,35	19,98
Stock Exchange	16,33	16,36	16,48	14,74	15,07	15,40	14,69	14,20	13,58	14,12	13,59	13,95	12,44
Telecommuncations	16,22	17,82	16,46	16,63	16,55	16,48	16,72	16,76	16,46	18,30	16,21	16,03	15,55
Textile Industries	15,89	15,75	15,78	15,61	15,24	15,28	15,39	15,43	15,42	17,44	15,21	15,10	15,09
Toys and Leisure	16,76	17,58	17,64	17,94	17,44	16,67	15,29	15,81	16,99	16,08	16,03	15,60	15,51
Transport and Logistics	17,14	16,54	16,46	16,38	16,11	16,12	16,15	16,03	15,83	16,79	16,00	15,90	15,65
Total	**15,71**	**15,90**	**15,70**	**15,49**	**15,00**	**14,94**	**14,86**	**14,89**	**15,16**	**14,88**	**14,63**	**14,58**	**14,18**

Table 20 – Average score, per year, per sector - Gunning-Fog Index (Holdings)

Sector	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Holding Company - Agriculture	16,64	16,47	17,70	16,54	15,22	14,37	14,46	15,15	14,99	15,95	16,22	16,15	17,58
Holding Company - Civil Construction	16,16	15,51	15,17	15,12	15,15	14,53	15,15	16,01	14,83	15,72	15,63	14,87	14,49
Holding Company - Commerce	17,05	17,35	17,46	16,75	16,49	16,47	16,34	16,90	16,44	16,75	16,84	16,74	15,79
Holding Company - Communication and Informatics	0,00	0,00	11,53	21,04	16,02	14,59	14,68	15,11	15,13	15,17	13,30	16,17	15,84
Holding Company - Education	16,38	16,68	16,87	15,63	17,17	16,68	16,69	17,05	17,01	16,31	16,77	15,53	15,21
Holding Company - Electricity	17,37	20,18	18,36	17,01	16,83	16,45	16,68	17,14	16,05	17,11	16,67	16,14	15,91
Holding Company - Financial Intermediary	17,29	16,68	16,50	16,77	16,84	24,53	16,17	17,18	21,42	16,16	17,35	18,01	17,74
Holding Company - Food	15,11	14,94	13,91	13,18	12,78	13,04	13,21	13,13	13,33	13,66	13,36	12,44	13,56
Holding Company - Graphical Design and Publishing	19,70	17,85	17,87	22,73	22,34	22,90	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Holding Company - Hospitality and Tourism	15,04	16,20	15,39	12,98	11,02	11,06	18,17	18,18	17,69	18,19	17,92	18,03	17,89
Holding Company - Insurance	15,00	14,39	14,82	13,95	19,84	12,79	12,26	13,01	12,84	14,92	15,39	14,69	14,90
Holding Company - Leasing	16,49	15,92	16,55	15,97	16,52	17,36	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Holding Company - Machines, Equipment, Vehicles and Parts	17,29	16,51	16,58	16,31	15,89	15,55	15,53	15,85	14,98	16,40	16,06	14,99	14,94
Holding Company - Medical Services	21,43	27,19	22,54	22,07	21,56	22,72	0,00	17,47	15,23	16,16	15,50	15,70	16,06
Holding Company - Metallurgy and Steel	15,96	15,81	15,95	16,17	15,82	15,29	15,35	15,95	15,42	15,92	15,66	15,55	15,15
Holding Company - Mineral Extraction	17,10	17,84	17,27	16,88	16,82	16,18	16,15	16,64	18,78	19,27	19,27	18,55	17,53
Holding Company - No Main Sector	18,91	19,27	18,35	18,17	18,32	18,20	18,21	18,33	18,27	19,98	18,13	18,19	18,08
Holding Company - Oil and Gas	18,31	18,17	19,09	19,34	18,69	16,64	16,13	16,18	16,18	16,26	15,83	15,52	15,88
Holding Company - Petrochemicals and Rubber	15,78	15,24	14,90	15,48	14,72	14,28	14,68	15,64	13,95	15,13	14,78	14,42	14,88
Holding Company - Pharmaceuticals and Hygiene	0,00	22,34	20,95	0,00	0,00	0,00	16,15	15,76	16,88	17,59	0,00	0,00	0,00
Holding Company - Pulp and Paper	15,22	14,37	13,55	12,03	11,99	12,24	13,04	13,46	13,82	14,61	14,80	14,98	14,74
Holding Company - Real Estate Credit	17,55	17,99	18,41	19,39	19,03	18,32	18,58	19,78	19,05	20,13	19,32	17,93	18,50
Holding Company - Sanitization and Utilities	16,57	16,44	16,19	16,52	14,22	14,23	13,66	14,33	13,14	14,13	13,56	13,61	13,71
Holding Company - Securities	20,02	21,22	21,49	20,57	19,75	19,48	19,15	21,89	20,57	21,17	22,04	22,24	22,63
Holding Company - Stock Exchange	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Holding Company - Telecommunications	18,03	17,76	18,03	18,30	18,10	18,00	18,63	18,84	17,08	17,23	16,73	16,78	16,75
Holding Company - Textile Industries	15,83	16,24	16,56	16,32	15,93	15,86	16,04	16,35	15,47	18,05	15,88	15,35	15,79
Holding Company - Toys and Leisure	20,45	20,48	20,15	16,32	17,16	16,24	20,60	15,93	16,18	21,99	20,22	16,32	21,34
Holding Company - Transport and Logistics	17,00	17,16	16,98	16,43	16,43	16,27	16,43	16,85	16,41	16,36	16,38	15,79	15,50
Total	**15,44**	**16,42**	**16,52**	**15,79**	**15,54**	**15,32**	**13,87**	**14,76**	**14,52**	**15,18**	**14,26**	**13,95**	**14,15**

Table 21 – Average score, per year, per sector - SMOG Index

Sector	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**	**2023**
Agriculture	14,67	13,46	12,05	12,17	11,52	12,07	12,27	12,14	12,52	12,29	12,40	12,62	12,61
Banking	12,70	12,74	12,91	12,93	12,77	12,58	12,66	12,86	12,81	13,21	13,16	13,34	13,18
Civil Construction	13,60	13,71	13,60	13,39	13,12	13,19	13,30	13,34	13,25	13,73	13,62	13,12	12,88
Commerce	13,90	13,56	13,36	13,03	12,80	12,70	12,81	13,08	12,93	13,15	13,09	12,88	12,81
Communication and Informatics	14,61	13,57	13,12	13,18	12,81	12,83	12,73	12,06	12,02	12,84	13,08	13,10	12,94
Drinks and Tabacco	14,10	13,83	15,06	11,71	11,19	11,11	11,12	11,77	11,56	11,60	11,52	12,24	12,46
Education	15,61	16,12	15,11	14,89	15,33	15,15	14,39	14,02	13,39	13,88	13,53	13,21	12,78
Electricity	13,28	13,32	13,23	13,02	12,95	12,80	12,98	13,20	13,07	13,73	13,75	13,50	13,32
Factoring	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Financial Intermediary	14,69	14,97	15,17	14,96	15,11	14,89	15,55	15,50	16,14	16,58	15,70	15,43	15,83
Food	13,73	13,87	13,75	13,71	13,55	13,49	13,13	13,13	13,03	13,22	13,12	12,48	12,43
Graphical Design and Publishing	13,62	12,90	13,15	12,97	12,83	12,84	12,05	12,35	12,30	0,00	0,00	0,00	0,00
Hospitality and Tourism	15,15	15,11	15,04	13,37	13,58	12,79	12,70	12,46	12,41	12,68	12,79	12,17	12,83
Insurance	14,08	13,68	15,12	14,46	13,78	13,28	12,55	12,36	12,21	12,75	12,86	13,35	13,06
Leasing	13,96	13,97	14,09	14,55	14,76	14,69	14,47	14,35	14,16	14,73	15,91	15,56	14,66
Machines, Equipment, Vehicles and Parts	13,33	13,28	13,31	13,34	13,21	13,06	13,17	13,34	13,16	13,28	13,20	13,06	12,80
Medical Services	13,00	12,91	13,04	12,88	12,75	12,68	12,93	12,91	12,58	12,49	12,72	12,52	12,17
Metallurgy and Steel	13,53	13,55	13,41	13,32	13,21	13,10	13,11	13,18	12,80	13,20	13,00	13,25	13,18
Mineral Extraction	15,08	14,66	14,21	14,24	14,18	13,97	14,20	14,40	16,17	15,90	15,22	14,36	14,11
Oil and Gas	16,49	16,32	15,85	14,91	14,68	14,90	14,71	15,24	15,55	16,38	15,88	15,40	14,09
Packaging	13,36	13,49	13,74	13,67	13,60	14,21	14,75	15,10	14,86	14,79	15,20	14,27	13,64
Petrochemicals and Rubber	12,75	12,63	12,55	12,61	12,74	12,82	12,73	12,73	12,52	13,16	12,92	12,86	12,81
Pharmaceuticals and Hygiene	14,81	15,20	14,94	14,65	14,42	14,14	13,89	13,75	14,31	13,30	13,21	13,25	13,12
Pulp and Paper	13,91	14,30	13,58	12,95	12,83	13,11	12,55	12,84	13,45	13,58	13,24	12,57	11,66
Real Estate Credit	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Reforestation	14,54	15,98	15,14	15,71	15,61	16,04	16,31	16,66	15,89	15,63	15,49	15,19	14,82
Sanitization and Utilities	14,27	14,26	14,29	14,60	14,11	14,26	14,31	14,27	14,14	14,37	14,10	13,83	13,81
Securities	16,27	16,46	16,60	16,70	16,49	16,14	16,00	16,14	16,91	17,21	17,27	17,18	16,35
Stock Exchange	14,11	14,11	14,14	12,88	13,10	13,36	12,78	12,48	12,21	12,36	12,27	12,31	11,99
Telecommuncations	13,76	13,93	13,96	14,14	14,09	13,95	14,18	14,21	14,13	14,39	13,88	13,70	13,26
Textile Industries	13,54	13,44	13,44	13,33	13,11	13,13	13,20	13,20	13,28	13,41	13,04	13,00	13,01
Toys and Leisure	14,16	14,79	14,82	15,14	14,69	14,04	13,06	13,48	13,27	13,61	13,63	13,41	13,29
Transport and Logistics	14,45	14,00	13,92	13,86	13,70	13,68	13,69	13,66	13,52	13,73	13,72	13,50	13,40
Total	**13,30**	**13,28**	**13,20**	**12,95**	**12,81**	**12,76**	**12,67**	**12,73**	**12,74**	**12,58**	**12,50**	**12,32**	**12,10**

Table 22 – Average score, per year, per sector - SMOG Index (Holdings)

Sector	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Holding Company - Agriculture	14,06	14,09	14,90	14,11	12,96	12,24	12,53	12,86	12,68	13,33	13,53	13,54	14,61
Holding Company - Civil Construction	13,80	13,36	13,06	13,13	13,20	12,57	13,09	13,72	12,87	13,44	13,30	12,80	12,52
Holding Company - Commerce	14,57	14,88	14,88	14,41	14,31	14,29	14,23	14,61	14,22	14,36	14,36	14,27	13,73
Holding Company - Communication and Informatics	0,00	0,00	8,96	12,22	13,54	12,65	12,45	12,86	12,92	12,88	11,74	13,72	13,43
Holding Company - Education	13,93	14,07	14,18	13,39	14,42	14,14	14,05	14,29	14,27	13,88	14,16	13,17	12,96
Holding Company - Electricity	14,17	14,44	14,33	14,31	14,26	14,00	14,27	14,58	13,73	14,55	14,16	13,85	13,59
Holding Company - Financial Intermediary	14,75	14,14	13,99	14,19	14,27	13,96	14,28	14,54	14,51	14,29	14,98	15,30	15,07
Holding Company - Food	12,82	12,76	12,01	11,56	11,34	11,49	11,53	11,49	11,55	11,86	11,65	11,30	11,71
Holding Company - Graphical Design and Publishing	16,57	15,24	15,16	18,97	18,62	19,12	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Holding Company - Hospitality and Tourism	13,14	13,81	13,15	11,73	10,44	10,45	15,25	15,16	14,71	15,21	15,03	15,00	14,88
Holding Company - Insurance	12,98	12,54	12,64	12,11	11,89	11,30	10,96	11,41	11,33	12,60	13,03	12,56	12,55
Holding Company - Leasing	13,83	13,37	13,76	13,23	13,75	14,40	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Holding Company - Machines, Equipment, Vehicles and Parts	14,57	13,99	13,97	13,74	13,47	13,21	13,25	13,53	12,96	14,01	13,69	12,81	12,82
Holding Company - Medical Services	18,43	15,84	19,20	18,84	18,52	19,35	0,00	15,19	13,34	13,79	13,24	12,10	13,26
Holding Company - Metallurgy and Steel	13,72	13,59	13,65	13,84	13,53	13,36	13,41	13,85	13,42	13,70	13,41	13,33	13,25
Holding Company - Mineral Extraction	14,53	15,02	14,53	14,32	14,27	13,80	13,80	14,12	15,77	16,21	16,28	15,79	14,89
Holding Company - No Main Sector	15,79	15,94	15,39	15,28	15,43	15,34	15,40	15,52	15,49	15,46	15,35	15,07	15,32
Holding Company - Oil and Gas	15,43	15,29	16,15	16,32	15,79	14,14	14,03	14,11	14,06	14,03	13,80	13,40	13,77
Holding Company - Petrochemicals and Rubber	13,36	13,04	12,72	13,19	12,66	12,29	12,65	13,30	12,08	12,84	12,66	12,36	12,62
Holding Company - Pharmaceuticals and Hygiene	0,00	18,60	17,53	0,00	0,00	0,00	13,72	13,31	14,26	14,86	0,00	0,00	0,00
Holding Company - Pulp and Paper	13,27	12,66	12,15	11,06	10,96	11,08	11,58	11,77	12,02	12,54	12,73	12,86	12,73
Holding Company - Real Estate Credit	14,95	15,02	15,32	16,14	15,88	15,41	15,59	16,48	15,95	16,69	16,10	15,22	15,46
Holding Company - Sanitization and Utilities	13,87	13,77	13,67	13,88	12,21	12,33	11,94	12,36	11,63	12,42	11,97	12,00	12,09
Holding Company - Securities	16,68	17,77	18,00	17,29	16,76	16,45	16,26	17,30	17,31	17,97	18,43	18,51	18,97
Holding Company - Stock Exchange	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Holding Company - Telecommunications	15,16	14,99	15,20	15,45	15,25	15,16	15,65	15,86	14,41	14,69	14,28	13,83	14,17
Holding Company - Textile Industries	13,49	13,79	14,03	13,86	13,69	13,66	13,64	13,99	13,25	12,60	13,38	13,07	13,58
Holding Company - Toys and Leisure	17,08	17,12	16,95	13,84	14,67	13,91	13,96	13,78	13,99	18,63	17,09	14,49	18,13
Holding Company - Transport and Logistics	14,43	14,57	14,45	14,04	13,82	13,92	14,12	14,39	14,05	13,94	13,96	13,48	13,34
Total	**13,08**	**13,57**	**13,93**	**13,26**	**13,10**	**12,90**	**11,78**	**12,57**	**12,30**	**12,79**	**12,15**	**11,86**	**12,05**

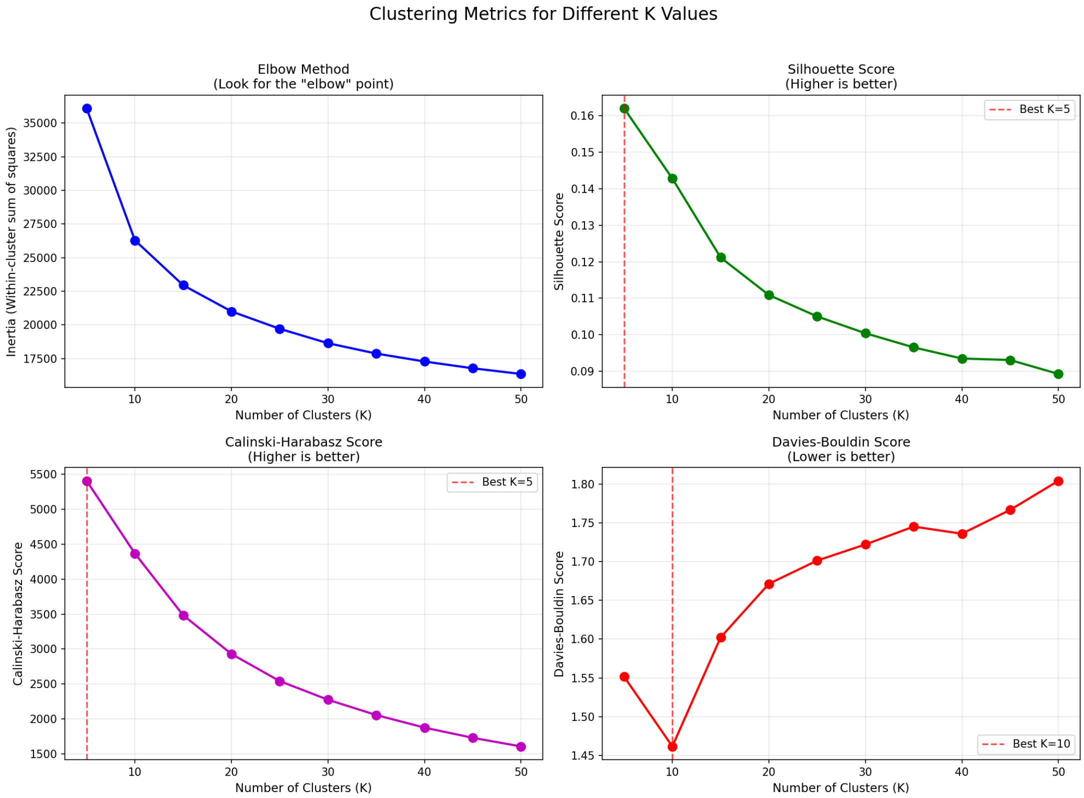
Table 23 – Average score, per year, per sector - Loughran-McDonald Index

Sector	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**	**2023**
Agriculture	11,70	12,06	12,14	12,00	11,88	11,90	11,86	11,88	11,90	11,99	12,01	12,13	12,15
Banking	12,12	12,16	12,22	12,27	12,29	12,33	12,31	12,31	12,32	12,14	12,23	12,40	12,47
Civil Construction	11,64	11,71	11,75	11,76	11,80	11,80	11,85	11,93	11,92	11,88	11,80	11,85	11,88
Commerce	11,93	11,90	11,92	11,94	11,89	11,92	11,87	11,94	11,93	12,07	12,00	12,00	12,03
Communication and Informatics	11,52	11,70	11,76	11,79	11,82	11,86	11,81	11,90	12,08	11,89	12,00	11,98	11,94
Drinks and Tabacco	12,19	12,26	12,32	12,51	12,46	12,68	12,71	12,67	12,83	12,78	12,74	12,54	12,42
Education	10,64	10,69	11,40	11,53	11,60	11,61	11,74	11,76	11,82	11,96	11,93	11,88	11,79
Electricity	12,00	12,03	12,05	12,10	12,13	12,15	12,16	12,12	12,13	12,08	12,06	12,03	11,95
Factoring	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Financial Intermediary	11,53	11,60	11,68	11,55	11,59	11,55	11,39	11,45	11,30	11,37	11,66	11,81	11,63
Food	11,70	11,94	11,92	11,98	11,97	12,04	11,91	11,91	11,94	11,95	12,00	12,03	12,12
Graphical Design and Publishing	12,07	12,09	12,23	12,22	12,18	12,11	11,89	12,11	12,00	0,00	0,00	0,00	0,00
Hospitality and Tourism	10,78	11,03	11,20	11,62	11,39	11,39	11,41	11,65	11,71	11,72	11,59	11,55	11,47
Insurance	12,28	12,11	11,61	11,63	11,76	11,83	12,15	12,35	12,42	12,26	12,31	12,33	12,35
Leasing	11,30	11,36	11,40	11,46	11,47	11,48	11,41	11,37	11,36	11,27	11,06	10,98	10,87
Machines, Equipment, Vehicles and Parts	11,75	11,74	11,85	11,81	11,83	11,77	11,86	11,86	11,90	11,93	11,96	11,95	11,98
Medical Services	11,95	11,97	12,14	12,13	12,12	12,02	12,13	12,08	12,16	12,19	12,13	12,16	12,14
Metallurgy and Steel	11,43	11,50	11,55	11,59	11,65	11,65	11,66	11,68	11,74	11,73	11,69	11,76	11,79
Mineral Extraction	11,01	11,16	11,32	11,32	11,18	11,12	11,04	11,00	10,82	11,33	11,51	11,74	11,60
Oil and Gas	11,11	10,75	10,80	10,61	10,59	10,91	11,12	11,10	10,77	11,27	11,34	11,41	11,44
Packaging	11,69	11,74	11,77	11,85	11,82	11,89	12,12	11,98	11,98	12,01	11,90	11,51	11,29
Petrochemicals and Rubber	11,97	11,82	11,72	11,89	11,86	11,87	11,78	11,83	11,78	11,83	11,80	11,91	11,91
Pharmaceuticals and Hygiene	11,69	11,71	11,59	11,82	11,89	11,76	11,82	11,73	11,56	11,68	11,77	11,73	11,78
Pulp and Paper	11,74	11,73	11,83	11,90	11,84	11,82	11,89	11,94	11,76	11,87	11,90	12,06	12,02
Real Estate Credit	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Reforestation	10,48	10,58	10,60	10,53	10,49	10,52	10,55	10,70	10,57	10,53	10,53	10,47	10,50
Sanitization and Utilities	11,91	11,88	11,92	11,94	11,89	11,90	11,93	11,93	11,94	11,96	12,01	11,98	11,93
Securities	10,72	10,88	10,94	10,97	10,90	10,89	10,86	10,89	10,85	10,79	10,70	10,83	10,93
Stock Exchange	12,40	12,35	12,37	12,11	12,23	12,25	12,48	12,32	12,35	12,23	12,29	12,29	12,43
Telecommuncations	11,66	11,57	11,57	11,58	11,58	11,63	11,66	11,71	11,73	11,71	11,72	11,76	11,85
Textile Industries	11,36	11,42	11,59	11,71	11,72	11,67	11,70	11,70	11,70	11,80	11,81	11,79	11,80
Toys and Leisure	11,36	11,31	11,29	11,33	11,24	11,10	11,31	11,45	11,57	11,52	11,77	11,76	11,76
Transport and Logistics	11,76	11,65	11,70	11,73	11,70	11,61	11,61	11,67	11,65	11,69	11,68	11,67	11,64
Total	**10,89**	**10,92**	**10,97**	**11,01**	**10,99**	**11,00**	**11,03**	**11,06**	**11,05**	**10,71**	**10,72**	**10,74**	**10,72**

Table 24 – Average score, per year, per sector - Loughran-McDonald Index (Holdings)

Sector	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**	**2023**
Holding Company - Agriculture	11,80	11,47	12,26	12,37	11,96	12,09	12,00	12,08	12,17	12,20	11,87	11,61	12,01
Holding Company - Civil Construction	11,87	11,92	12,16	11,94	11,97	12,06	11,67	11,84	11,97	11,92	11,86	11,90	11,97
Holding Company - Commerce	11,38	11,46	11,64	11,31	11,39	11,37	11,46	11,57	11,64	11,82	11,77	11,86	11,92
Holding Company - Communication and Informatics	0,00	0,00	11,52	11,81	11,48	11,66	11,74	11,68	11,89	11,87	11,27	11,69	11,59
Holding Company - Education	11,75	11,41	11,33	11,48	11,13	11,16	11,15	11,23	11,29	11,28	11,29	11,47	11,44
Holding Company - Electricity	11,79	11,71	11,61	11,66	11,68	11,66	11,69	11,65	11,72	11,66	11,74	11,66	11,60
Holding Company - Financial Intermediary	11,45	11,54	11,72	11,62	11,53	11,80	11,56	11,69	11,78	11,69	11,51	11,68	11,77
Holding Company - Food	11,98	11,99	11,96	11,96	11,97	12,06	12,02	12,08	12,05	12,03	12,02	12,02	11,91
Holding Company - Graphical Design and Publishing	11,26	11,05	10,63	10,97	10,72	10,89	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Holding Company - Hospitality and Tourism	11,46	11,27	11,47	11,83	11,66	11,55	10,98	11,19	11,31	11,45	11,49	11,45	11,50
Holding Company - Insurance	12,19	12,30	12,23	12,06	12,20	12,28	12,47	12,51	12,37	12,40	12,46	12,27	12,27
Holding Company - Leasing	11,06	10,96	10,49	10,34	10,35	10,34	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Holding Company - Machines, Equipment, Vehicles and Parts	11,42	11,52	11,61	11,78	11,67	11,64	11,61	11,89	12,01	12,03	11,69	11,83	11,56
Holding Company - Medical Services	12,71	13,11	12,72	12,57	12,40	12,24	0,00	12,02	12,38	12,40	12,36	12,46	12,21
Holding Company - Metallurgy and Steel	11,88	11,80	11,92	11,88	11,76	11,79	11,81	11,94	11,82	11,81	11,87	11,89	11,85
Holding Company - Mineral Extraction	11,55	11,37	11,30	11,41	11,37	11,47	11,55	11,54	11,52	11,50	11,42	11,41	11,39
Holding Company - No Main Sector	10,91	11,00	10,97	11,01	10,91	10,95	11,02	11,04	11,09	11,24	11,23	11,26	11,25
Holding Company - Oil and Gas	11,05	11,16	11,29	11,31	11,52	11,73	11,89	11,85	11,84	12,01	12,13	12,19	12,13
Holding Company - Petrochemicals and Rubber	11,91	12,02	12,12	12,11	12,36	12,60	12,76	12,92	12,83	12,78	12,58	12,23	12,54
Holding Company - Pharmaceuticals and Hygiene	0,00	10,81	10,63	0,00	0,00	0,00	11,70	12,02	11,34	11,84	0,00	0,00	0,00
Holding Company - Pulp and Paper	12,10	12,37	12,30	12,21	12,21	12,24	12,19	12,27	12,34	12,46	12,44	12,43	12,37
Holding Company - Real Estate Credit	11,36	11,27	10,98	10,88	10,72	10,77	10,92	11,09	10,91	11,03	10,90	10,87	11,03
Holding Company - Sanitization and Utilities	11,88	11,91	11,88	11,66	11,92	11,98	12,10	12,20	12,22	12,20	12,01	11,90	12,00
Holding Company - Securities	10,37	10,60	10,96	10,94	10,96	11,18	11,13	10,96	10,65	10,67	10,79	10,97	11,23
Holding Company - Stock Exchange	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Holding Company - Telecommunications	11,43	11,28	11,27	11,43	11,29	11,15	11,09	11,12	11,57	11,68	11,73	11,87	11,88
Holding Company - Textile Industries	11,73	11,88	11,74	11,78	11,84	11,66	11,80	11,83	11,12	11,42	11,57	11,88	11,81
Holding Company - Toys and Leisure	10,72	10,69	10,57	11,50	11,51	11,55	11,90	11,12	11,10	11,58	11,63	11,32	11,55
Holding Company - Transport and Logistics	11,52	11,43	11,39	11,47	11,57	11,55	11,66	11,68	11,74	11,86	11,97	11,91	11,93
Total	**10,36**	**10,73**	**11,13**	**10,80**	**10,76**	**10,81**	**10,06**	**10,52**	**10,51**	**10,58**	**10,12**	**10,14**	**10,16**

APPENDIX B – Additional Tests Figures



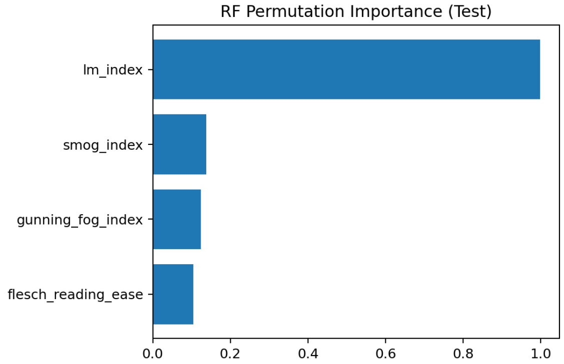


Figure 3 – Random Forest MDA Feature Importance

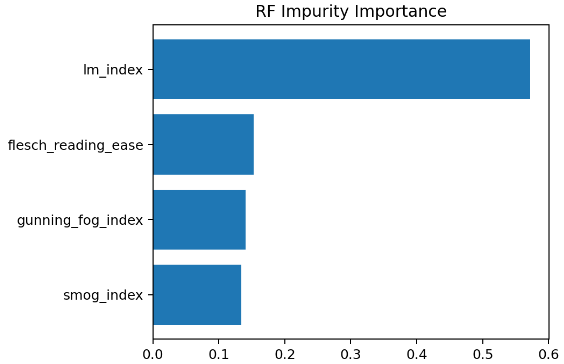


Figure 4 – Random Forest MDI Feature Importance

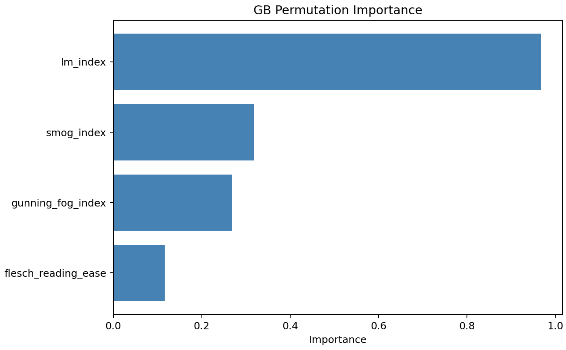


Figure 5 – Gradient Boosting MDA Feature Importance

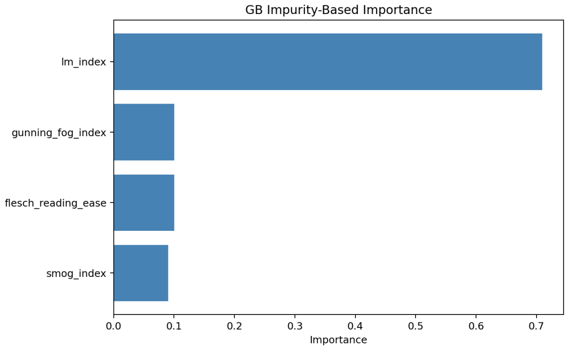


Figure 6 – Gradient Boosting MDI Feature Importance