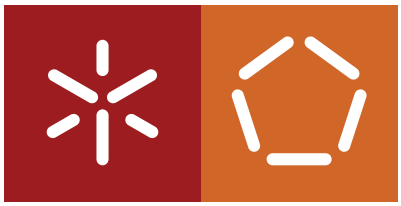
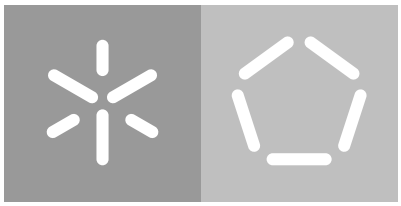


University of Minho
School of Engineering
Department of Informatics

Marcelo Queirós Pinto

Combining Data and Text Mining techniques for automatic analysis of Financial Reports

October 2019



University of Minho
School of Engineering
Department of Informatics

Marcelo Queirós Pinto

Combining Data and Text Mining techniques for automatic analysis of Financial Reports

Master dissertation
Master Degree in Computer Science and Engineering

Dissertation supervised by
Professor Doctor Paulo Cortez
Professor Doctor Nelson Areal

October 2019

ACKNOWLEDGEMENTS

I consider that the personal stability is fundamental to professional success, and in this sense, I give my deep gratitude to the people I can always count on in my personal life, and who were undoubtedly fundamental throughout this process. Thank you for listening and giving feedback, for enduring me, for motivating me even on the toughest days and essentially for all your concern.

I would like to thank Professor Paulo Cortez and Professor Nelson Areal for the fantastic orientation of the project, the meetings, the suggestions and especially for all the learning I have gathered. Both were two of the best professors I have ever dealt with.

Since during this thesis I worked in 2 companies, I want to thank them for all the flexibility and support they provided whenever I needed it. They have given me tremendous technical knowledge important for this thesis as well as important personal skills for my future. The colleagues and friends I made in these companies were always available to give me any opinion on this thesis and I really appreciate it.

To my master and bachelor colleagues at the University of Minho and University of Trás-os-Montes and Alto-Douro, who at the same time many of them became personal friends, I greatly appreciate all your willingness to help me and give good opinions.

Marcelo Queirós

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

ABSTRACT

The application of Data Science techniques, specifically Natural Language Processing (NLP) and Machine Learning, in financial markets is of immense interest to investors, as these techniques can have a potential economic impact. In particular, stock markets represent an opportunity that has been exploited in several ways, such as using market opinions (e.g., news, blogs) to predict the direction of price movement or even volatility.

This study analyses the 10-K documents of the S&P 100 index for 10 years (2008-2017), which contains the 102 largest companies in the United States of America. The 10-K is an annual financial report required by the United States Securities and Exchange Commission (SEC), which describes the financial performance of a company. Recent research suggests that the readability of a company's 10-K text document may influence its future financial performance, since the way the market perceives textual information also depends on the readability of that text. In this sense, this work aims to understand the relationship between 48 readability metrics applied to these reports and the corresponding future financial performance of these companies. A clustering approach was applied over these readability metrics, aiming to identify distinct and valuable readability clusters. As an external evaluation, we assessed the information value of the clusters by analyzing 3 future crash risk metrics, that are often used to assess the companies' financial performance.

Keywords: Data Science; Financial performance prediction; Natural Language Processing; Readability evaluation; Stock markets.

RESUMO

A aplicação das técnicas de Ciência de Dados, especificamente Processamento de Linguagem Natural e *Machine Learning*, nos mercados financeiros é de imenso interesse para os investidores, uma vez que podem ter um potencial impacto económico. Em particular, os mercados de ações representam uma oportunidade que tem sido explorada de várias formas, como no uso de informações de mercado (por exemplo notícias, blogs) para prever a direção do movimento dos preços ou mesmo o movimento da volatilidade.

Este estudo analisa os documentos 10-K do índice S&P 100 durante 10 anos (2008-2017), que contém as 102 maiores empresas dos Estados Unidos da América. O 10-K é um relatório financeiro anual exigido pela Comissão de Valores Mobiliários dos Estados Unidos (SEC), que descreve o desempenho financeiro de uma empresa. Pesquisas recentes sugerem que a legibilidade do documento de texto 10-K de uma empresa pode influenciar o seu desempenho financeiro futuro, uma vez que a forma como o mercado perceciona as informações textuais também depende da legibilidade desse texto. Neste sentido, este trabalho visa compreender a relação entre 48 métricas de legibilidade aplicadas a esses relatórios e o desempenho financeiro futuro correspondente dessas empresas. Uma abordagem de agrupamento de dados foi aplicada nestas métricas de legibilidade, com o objetivo de identificar grupos de legibilidade distintos e relevantes. Com uma avaliação externa, avaliamos o valor das informações desses grupos analisando três métricas de *crash risk* futuro, que são frequentemente usadas para avaliar o desempenho financeiro das empresas.

Palavras-Chave: Avaliação da legibilidade; Ciência de Dados; Mercado de ações; Previsão do desempenho financeiro; Processamento de Linguagem Natural.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS LIST

AD	Average Distance
ADM	Average Distance between Means
APN	Average Proportion of Non-overlap
CIK	Central Index Key
CRISP-DM	Cross-Industry Standard Process for Data Mining
CRSP	Center for Research in Security Prices
DUVOL	Down-to-Up Volatility
FOM	Figure Of Merit
NCSKEW	Negative Coefficient of Skewness
NLP	Natural Language Processing
PCA	Principal Component Analysis
SEC	United States Securities and Exchange Commission
SIC	Standard Industrial Classification
SVM	Support Vector Machines

1

INTRODUCTION

1.1 CONTEXT AND MOTIVATION

The study by Gantz and Reinsel [2012] points out that from 2005 to 2020, the digital universe will grow 300 times, from 130 exabytes to 40000 exabytes and the average of data generated for every person in the world is about 5200 gigabytes, considering man, woman, and child. From 2012 until 2020, the digital universe will double every two years, representing 16 times more data in 2020 than in 2012. The reasons for this growth of data are various, including the growth of the web and internet of things and the massive use of information technologies, generating a huge amount of data, giving rise to the new term known as Big Data. Under this context, Data Mining, using the power of Machine Learning, has an enormous potential, allowing the extraction of useful knowledge from raw data.

In particular, the Data Mining subarea, in which the raw material is the text, known as Text Mining, is growing substantially. The evolution of Text Mining, and in particular in Natural Language Processing (NLP) can greatly impact business, since most of the human information available is effectively text. Other areas impacted also include management, economics and finance. Investors are particularly interested in how new technologies can help them invest. There are already some methodologies that use Text Mining and NLP in financial domains, for example, by using market news to predict the direction of the price movement or even the volatility movement [Atkins et al., 2018].

Thus, the application of NLP and Machine Learning in financial markets is of immense interest to investors, and these technologies can have a economic impact. Recent research, such argued by Li [2008], Lawrence [2013] and Yu and Miller [2010], points to a possible relationship between how readable a financial report is and future financial results, which serves as the foundational motivation for this dissertation work: to understand the relationship between the readability of financial reports and the

companies' future crash risk. Future crash risk metrics are often used to assess the companies' financial performance. Realizing the relationship between present business components such as financial reporting readability, it is possible to build a crash risk forecasting model using readability indicators.

1.2 OBJECTIVES

As Li [2008], Lawrence [2013] and Yu and Miller [2010] research indicates, it is possible that readability measures can reveal insights about the current and future financial state of companies. These studies indicate, for example, that financial report texts can reveal components as the information asymmetry, the phenomenon that occurs when two or more economic agents hold different qualitative or quantitative information, defined in microeconomics as market failure, as well as it can describe the uncertainty of the information environment and the financial stress of companies.

In this particular study, annual financial reports (10-K) will be analysed, concluding their readability from 48 different readability metrics (within our knowledge, there is no other study that analyses such large amount of readability measures). Afterwards, this work aims to evaluate the relationship of the readability of these reports and the future crash risk, obtained from the three measures. Readability metrics measure how readable the financial reports are and the future crash risk allows to evaluate the future company's performance. In this project, future crash risk is obtained from the day following the publication of the financial report until the publication of the next report (next year). This study is intended to contribute to the increase of knowledge about the mentioned relationship. The following are the particular set of objectives that should be achieved. Table 1 details the designed plan.

1. Business understanding: study the different readability and future crash risk metrics by reviewing their literature.

2. Data preparation: Financial reports will be obtained from US companies using the SEC EDGAR platform and financial information (used to calculate future crash risk metrics) will be obtained from the Center for Research in Security Prices (CRSP) databases and Professor K. French website [Kenneth, 2019].

3. Analyse all S&P 100 index companies, for 10 years.

4. Obtain 48 readability measures from this financial reports, analysing different components of their text.

5. Obtain future crash risk measures from the day after the financial report is published until the day before the publication of the next financial report (in the next year).

6. Analyse the relationship between readability measures and future crash risk measures of the same company for the same year (relate the readability of a report to next year's financial performance, starting the day after the financial report is published).

The following is the plan to achieve the objectives of this work:

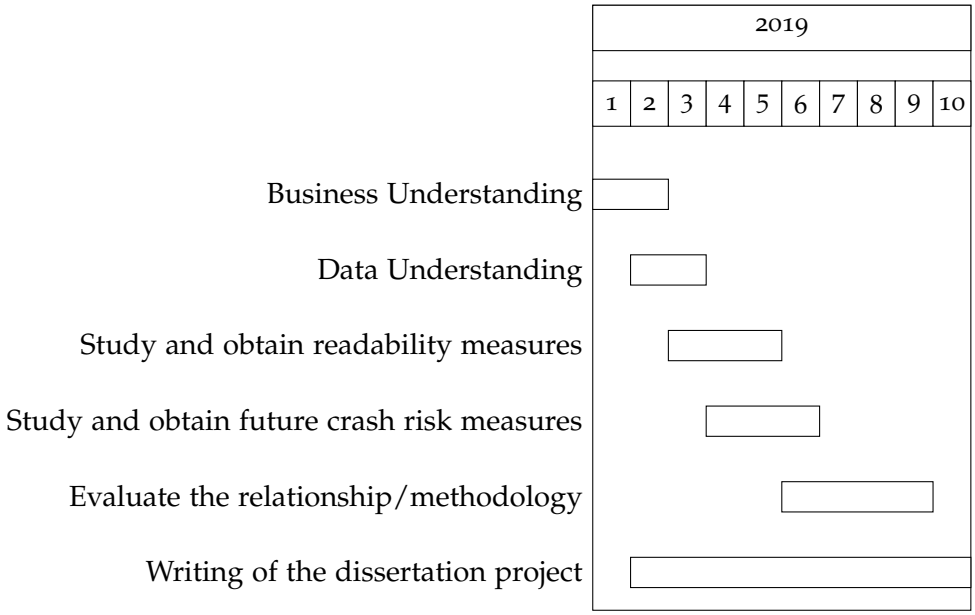


Table 1: Objectives in a Gantt chart.

1.3 DOCUMENT ORGANIZATION

This document is organized in terms of the chapters:

1. Chapter 2: State of the art, which organizes and exposes the research found as well as it is intended to explain all 48 readability measures and their components, also showing their calculation formula, as seen in Section 2.2.4. It also refers and explains future crash risk measures and how financial information is used in their formulas. This section analyses recent research that demonstrates evidences of how readability can be related to financial components.

2. Chapter 3: Development. This phase goes thorough the development process, explaining the necessary steps to analyses the relationship between the readability and future crash risk. It includes business understanding, data understanding, data preparation as well as the finally analysis, to evaluate the relationship, in terms of technical issues.

3. Chapter 4: Experiments, which is intended to evaluate the mentioned relationship, making it possible to draw conclusions about the contribution of this project in the financial and Data Science fields.

4. Chapter 5: Conclusions, which presents the main conclusions of the study and the various directions for future work.

2

STATE OF THE ART

This chapter presents the main concepts approached in this work, also disclosing the relevant works in this domain.

2.1 DEFINING READABILITY

There is no universal definition of readability [Loughran and McDonald, 2014]. Some definitions focus solely on the style of writing, content, coherence and organization. For example: Klare et al. [1963] mentions it is "the ease of understanding or comprehension due to the style of writing". This definition focuses the majority on the clarity, transparency and accessibility of the text as well as the ease with which the readers can understand the text. On the other hand, Mc Laughlin [1969] and DuBay [2004] define readability as "the degree to which a given class of people finds a reading attractive and understandable" focusing on the public-target, insofar as readability takes into account the characteristics of the readers and their limitations. Davison and Kantor [1982] emphasize that "background knowledge assumed in the reader" is more important than "trying to make a text fit a readability level defined by a formula".

2.2 READABILITY MEASURES

A readability measure aims to probabilistically measure how readable a text is. It allows to know the readability of a text and thus be aware of its complexity. This section will cover dozens of redability measures, describing them and discussing how they are calculated. In this project, a total of 48 readability measures were used to increase the efficiency of the approach.

2.2.1 *Word count and file size*

Some measures are based on the notion of overwriting: when documents are written with too much detail, using dispensable and superfluous words that makes the document too long and consequently difficult to process and understand, influencing readability [Bonsall IV et al., 2017]. Therefore, the use of attributes such as the number of words in the document [You and Zhang, 2009] and the file size, (e.g., number of megabytes) are examples of overwriting measures [Loughran and McDonald, 2014]. These measures may be inefficient because they capture other factors that do not affect readability: using the file size may consider parts of the document that are independent of financial information and that vary in all financial reports (e.g., pictures, bond indentures, compensation, supplier/customer agreements) [Loughran and McDonald, 2014]. Furthermore, these measures will be associated with a lack of interpretation of semantics, which impairs readability.

2.2.2 *Fog Index*

Gunning [1952] developed the Fog Index, considered by many researchers as a primary measure of readability, since it has a good measure of difficult complexity text and is easy to calculate and adaptable to the computational level through the Fog Index formula. The Fog Index has 2 elements: the average sentence size and the percentage of complex words (i.e., words with three or more syllables). The final measure will be a sum of 2 elements multiplied by a scalar (0.4).

$$0.4 \left[\frac{words}{sentences} + 100 \left(\frac{complexWords}{words} \right) \right] \tag{1}$$

Fog Index is a good sign of hard-to-read text but it has limitations: There are words that have three or more syllables (definition for complex words), which are not really difficult. For example, ”interesting” is not generally thought to be a difficult word, although it has four syllables (then classified as a complex word). Also, a short word can be difficult even if it is not used very often by most people, like the word “bilk”. The frequency with which words are in normal use affects the readability of text [Seely, 2013].

2.2.3 *Bog Index*

Bonsall IV et al. [2017], recognizing the limitations of the Fog Index, developed a methodology capable of capturing the plain English attributes of disclosure, recommended by linguistics experts and highlighted in the SEC's Plain English Handbook [SEC, 1998]. An example of attributes to capture would be a sentence that contains "I made an application", which has a readability mistake termed by linguistics as a hidden verb, which should have been replaced by "I applied". Bog Index is computed as the sum of three multifaceted components:

$$\Big[BogIndex = SentenceBog + WordBog - Pep\Big] \tag{2}$$

1. Sentence Bog identifies subjects related to sentence size, such as the average sentence size, which is squared and scaled by a standard long sentence limit of 35 words per sentence.

2. Word Bog has 2 main subcomponents: (a) plain English style problems and (b) word difficulty. The calculation of Word Bog is the sum of (a) plain English style problems and (b) word difficulty multiplied by 250 and divided by the number of words. The (a) plain English style problems is a combination of issues highlighted in the SEC's Plain English Handbook [SEC, 1998]: passive verbs, hidden verbs, overwriting, legal terms, cliches, abstract words and wordy phrases. The (b) word difficulty is calculated based on the difficulty for general vocabulary (i.e., heavy words), abbreviations and specialist terms [Bonsall IV et al., 2017].

3. Pep identifies writing attributes that facilitate understanding of texts by readers, as interesting and easy words. Pep is calculated as the sum of this components multiplied by 25 (Word Bog are multiplied by 250) and scaled by the number of words in the document plus sentence variety (i.e., the standard deviation of sentence length multiplied by ten and scaled by the average sentence length) [Bonsall IV et al., 2017].

In contrast to the Fog Index, which considers a word with three or more syllables, the Bog Index measures word difficulty using a proprietary list of over 200,000 words based on familiarity and assesses penalties between zero and four points based on

a combination of the word's familiarity and precision (abstract words receive higher scores).

2.2.4 *All 48 readability measures used in this study*

For this study, 48 readability measures were used, which analyze several components of textual complexity. From Table 2 to Table 9, it is presented the 48 readability measures: a description about them, their calculation formula and the main bibliographic reference for the measure. These measures are not uniform, i.e., some measures suggest a higher readability when their values are higher, while others produce numbers in the opposite direction, with higher values indicating more difficult texts. Thus, we also present this information in the tables. Nevertheless, to facilitate the analysis of the results, all measures were rescaled in an uniform fashion, assuring that larger values are related with more readable documents. The rescaling was based on the simpler symmetrical transformation, where only the measures that are lower for more readable texts invert the signal. We adopted the following notation to define the readability measures:

1. n_w = number of words;
2. n_c = number of characters;
3. n_{st} = number of sentences;
4. n_{sy} = number of syllables;
5. n_{wf} = number of words matching the Dale-Chall List of 3000 ”familiar words”;
6. ASL = Average Sentence Length: number of words / number of sentences;
7. AWL = Average Word Length: number of characters / number of words;
8. AFW = Average Familiar Words: count of words matching the Dale-Chall list of 3000 ”familiar words” / number of all words; and
9. n_{wd} = number of ”difficult” words not matching the Dale-Chall list of ”familiar” words.

Measure name	Description	More readable when:
ARI	Automated Readability Index [Smith and Senter, 1967]. $0.5ASL + 4.71AWL - 21.34$	Lower
ARI.Simple	A simplified version of Smith and Senter [1967] Automated Readability Index. $ASL + 9AWL$	Lower
ARI.NRI	Automated Readability Index [Smith and Senter, 1967] with revised parameters from the Navy Readability Indexes. $0.4ASL + 6AWL - 27.4$	Lower
Bormuth.MC	Bormuth [1969] Mean Cloze Formula. $0.886593 - 0.03640 * AWL + 0.161911 * AFW - 0.21401 * ASL - 0.000577 * ASL^2 - 0.000005 * ASL^3$	Higher
Bormuth.GP	Bormuth [1969] Grade Placement score. $4.275 + 12.881M - 34.934M^2 + 20.388M^3 + 26.194CCS - 2.046CCS^2 - 11.767CCS^3 - 42.285(M * CCS) + 97.620(M * CCS)^2 - 59.538(M * CCS)^2$ where M is the Bormuth Mean Cloze Formula as in "Bormuth" above and CCS is the Cloze Criterion Score [Bormuth, 1968].	Lower
Coleman	Coleman [1971] Readability Formula 1. $1.29 * (100 * n_{wsy=1}/n_w) - 38.45$ where $n_{wsy=1}$ = the number of words with 1 syllable	Higher
Coleman.C2	Coleman [1971] Readability Formula 2. $1.16 * (100 * n_{wsy=1}/n_w) + 1.48 * (100 * n_{st}/n_w) - 37.95$	Higher

Table 2: All 48 readability measures used in this study - Part 1 of 8.

Measure name	Description	More readable when:
Coleman.Liau .ECP	Coleman-Liau Estimated Cloze Percent (ECP) [Coleman and Liau, 1975]. $141.8401 - (0.214590 * 100 * AWL) + (1.079812 * n_{st} * 100/n_w)$	Higher
Coleman.Liau .grade	Coleman-Liau Grade Level [Coleman and Liau, 1975]. $-27.4004 * Coleman.Liau.ECP/100 + 23.06395$	Lower
Coleman.Liau .short	Coleman-Liau Index [Coleman and Liau, 1975]. $5.88 * AWL + (0.296 * n_{st}/n_w) - 15.8$	Lower
Dale.Chall	The New Dale-Chall Readability formula [Chall and Dale, 1995]. $64 - (0.95 * 100 * n_{wd}/n_w) - (0.69 * ASL)$	Higher
Dale.Chall.old	The original Dale-Chall Readability formula [Dale and Chall, 1948]. $0.1579 * 100 * n_{wd}/n_w + 0.0496 * ASL[+3.6365]$ The additional constant 3.6365 is only added if $(n_{wd}/n_w) > 0.05$.	Lower
Dale.Chall.PSK	The Powers-Sumner-Kearl Variation of the Dale and Chall Readability formula [Powers et al., 1958]. $(0.1155 * 100 * n_{wd}/n_w) + (0.0596 * ASL) + 3.2672$	Lower
Danielson .Bryan	Danielson and Bryan [1963] Readability Measure 1. $(1.0364 * n_c/n_{blank}) + (0.0194 * n_c/n_{st}) - 0.6059$ where n_{blank} = the number of blanks.	Lower

Table 3: All 48 readability measures used in this study - Part 2 of 8.

Measure name	Description	More readable when:
Danielson .Bryan.2	Danielson and Bryan [1963] Readability Measure 2. $131.059 - (10.364 * n_c/n_{blank}) + (0.0194 * n_c/n_{st})$ where n_{blank} = the number of blanks.	Higher
Dickes.Steiwer	Dickes-Steiwer Index [Dickes and Steiwer, 1977]. $235.95993 - (73.021 * AWL) - (12.56438 * ASL) - (50.03293 * TTR)$ where TTR is the Type-Token Ratio.	Higher
DRP	Degrees of Reading Power Bormuth [1969]. $(1 - Bormuth.MC) * 100$ where Bormuth.MC refers to Bormuth [1969] Mean Cloze Formula.	Lower
ELF	Easy Listening Formula [Fang, 1966]. $n_{wsy>=2}/n_{st}$. where $n_{wsy>=2}$ the number of words with 2 syllables or more.	Lower
Farr.Jenkins .Paterson	Farr-Jenkins-Paterson's Simplification of Flesch's Reading Ease Score [Farr et al., 1951]. $-31.517 - (1.015 * ASL) + (1.599 * n_{wsy=1}/n_w)$ where $n_{wsy=1}$ = the number of one-syllable words	Higher.
Flesch	Flesch's Reading Ease Score [Flesch, 1948]. $206.835 - (1.015 * ASL) - (84.6 * (n_{sy}/n_w))$	Higher
Flesch.PSK	The Powers-Sumner-Kearl's Variation of Flesch Reading Ease Score [Powers et al., 1958]. $(0.0078 * ASL) + (4.55 * n_{sy}/n_w) - 2.2029$	Lower

Table 4: All 48 readability measures used in this study - Part 3 of 8.

Measure name	Description	More readable when:
Flesch.Kincaid	Flesch-Kincaid Readability Score [Kincaid et al., 1975]. $$0.39 * ASL + 11.8 * (n_{sy}/n_w) - 15.59$$	Lower
FOG	Gunning's Fog Index [Gunning, 1952]. $$0.4 * (ASL + 100 * (n_{wsy>=3}/n_w)$$ where $n_{wsy>=3}$ = the number of words with 3-syllables or more. The scaling by 100 arises because the original Fog Index is based on just a sample of 100 words).	Lower
FOG.PSK	The Powers-Sumner-Kearl Variation of Gunning's Fog Index [Powers et al., 1958]. $$3.0680 * (0.0877 * ASL) + (0.0984 * 100 * (n_{wsy>=3}/n_w)$$ where $n_{wsy>=3}$ = the number of words with 3-syllables or more. The scaling by 100 arises because the original Fog Index is based on just a sample of 100 words).	Lower
FOG.NRI	The Navy's Adaptation of Gunning's Fog Index [Kincaid et al., 1975]. $$(((n_{wsy<3} + 3 * n_{wsy=3})/(100 * n_{st}/n_w)) - 3)/2$$ where $n_{wsy<3}$ = the number of words with less than 3 syllables and $n_{wsy=3}$ = the number of 3-syllable words. The scaling by 100 arises because the original Fog Index is based on just a sample of 100 words).	Lower
FORCAST	FORCAST (Simplified Version of FORCAST.RGL) [Caylor and Sticht, 1973]. $$20 - (n_{wsy=1} * 150)/(n_w * 10)$$ where $n_{wsy=1}$ = the number of one-syllable words. The scaling by 150 arises because the original FORCAST index is based on just a sample of 150 words.	Lower

Table 5: All 48 readability measures used in this study - Part 4 of 8.

Measure name	Description	More readable when:
FORCAST.RGL	FORCAST.RGL [Caylor and Sticht, 1973]. $$20.43 - 0.11 * (n_{wsy=1} * 150)/(n_w * 10)$$ where $n_{wsy=1}$ = the number of one-syllable words. The scaling by 150 arises because the original FORCAST index is based on just a sample of 150 words.	Lower
Fucks	Fucks [1955] Stilcharakteristik (Style Characteristic). $$AWL * ASL$$	Lower
Linsear.Write	Linsear Write [Klare, 1974]. $$[(100 - (100 * n_{wsy<3}/n_w)) + (3 * 100 * n_{wsy>=3}/n_w)]/(100 * n_{st}/n_w)$$ where $n_{wsy<3}$ = the number of words with less than 3 syllables, and $n_{wsy>=3}$ = the number of words with 3-syllables or more. The scaling by 100 arises because the original Linsear. Write measure is based on just a sample of 100 words).	Lower
LIW	Björnsson [1968] Läs barhets index. $$ASL + (100 * n_{wsy>=7}/n_w)$$ where $n_{wsy>=7}$ = the number of words with 7-syllables or more. The scaling by 100 arises because the Läsbarhetsindex index is based on just a sample of 100 words).	Lower
nWS	Neue Wiener Sachtextformeln 1 [Bamberger and Vanecek, 1984]. $$(19.35 * n_{wsy>=3}/n_w) + (0.1672 * ASL) + (12.97 * n_{wchar>=6}/n_w) - (3.27 * n_{wsy=1}/n_w) - 0.875$$ where $n_{wsy>=3}$ = the number of words with 3 syllables or more, $n_{wchar>=6}$ = the number of words with 6 characters or more, and $n_{wsy=1}$ = the number of one-syllable words.	Lower

Table 6: All 48 readability measures used in this study - Part 5 of 8.

Measure name	Description	More readable when:
nWS.2	Neue Wiener Sachtextformeln 2 [Bamberger and Vanecek, 1984] $(20.07 * n_{wsy>=3}/n_w) + (0.1682 * ASL) + (13.73 * n_{wchar>=6}/n_w) - 2.779$ where $n_{wsy>=3}$ = the number of words with 3 syllables or more, and $n_{wchar>=6}$ = the number of words with 6 characters or more.	Lower
nWS.3	Neue Wiener Sachtextformeln 3 [Bamberger and Vanecek, 1984] $(29.63 * n_{wsy>=3}/n_w) + (0.1905 * ASL) - 1.1144$ where $n_{wsy>=3}$ = the number of words with 3 syllables or more.	Lower
nWS.4	Neue Wiener Sachtextformeln 4 [Bamberger and Vanecek, 1984] $(27.44 * n_{wsy>=3}/n_w) + (0.2656 * ASL) - 1.693$ where $n_{wsy>=3}$ = the number of words with 3 syllables or more.	Lower
RIX	Anderson [1983] Readability Index. $n_{wsy>=7}/n_{st}$ where $n_{wsy>=7}$ = the number of words with 7-syllables or more.	Lower
Scrabble	Scrabble is a game that contains a dictionary to give each letter a score, this score is used in this measure. Mean Scrabble letter values of all words Scrabble values are for English.	Lower

Table 7: All 48 readability measures used in this study - Part 6 of 8.

Measure name	Description	More readable when:
SMOG	Simple Measure of Gobbledygook (SMOG) [Mc Laughlin, 1969]. $1.043 * sqrt(n_{wsy>=3} * 30/n_{st}) + 3.1291$ where $n_{wsy>=3}$ = the number of words with 3 syllables or more. This measure is regression equation D in McLaughlin's original paper.	Lower
SMOG.C	SMOG (Regression Equation C) [Mc Laughlin, 1969]. $0.9986 * sqrt(n_{wsy>=3} * (30/n_{st}) + 5) + 2.8795$ where $n_{wsy>=3} = n_{wsy>=3}$ = the number of words with 3 syllables or more. This measure is regression equation C in McLaughlin's original paper.	Lower
SMOG.Simple	Simplified Version of Mc Laughlin [1969] SMOG Measure. $sqrt(n_{wsy>=3} * 30/n_{st}) + 3$	Lower
SMOG.de	Adaptation of Mc Laughlin [1969] SMOG Measure for German Texts. $sqrt(n_{wsy>=3} * 30/n_{st}) - 2$	Lower
Spache	Spache [1953] Readability Measure. $0.121 * ASL + 0.082 * n_{wnotinspache}/n_w) + 0.659$ where $n_{wnotinspache}$ = number of unique words not in the Spache word list.	Lower
Spache.old	Spache [1953] Readability Measure (Old). $0.141 * ASL + 0.086 * (n_{wnotinspache}/n_w) + 0.839$ where $n_{wnotinspache}$ = number of unique words not in the Spache word list.	Lower

Table 8: All 48 readability measures used in this study - Part 7 of 8.

Measure name	Description	More readable when:
Strain	Strain Index. $n_{wsy}/(n_{st}/3)/10$ The scaling by 3 arises because the original Strain index is based on just the first 3 sentences.	Lower
Traenkle.Bailer	Tränkle and Bailer [1984] Readability Measure 1. $224.6814 - (79.8304 * AWL) + (12.24032 * ASL) - (1.292857 * 100 * n_{prep}/n_w)$ where n_{prep} = the number of prepositions. The scaling by 100 arises because the original Tränkle & Bailer index is based on just a sample of 100 words.	Higher
Traenkle .Bailer2	Tränkle and Bailer [1984] Readability Measure 2. $234.1063 - 96.11069 * AWL - 2.05444 * 100 * (n_{prep}/n_w) - 1.02805 * 100 * (n_{conj}/n_w)$ where n_{prep} = the number of prepositions, n_{conj} = the number of conjunctions, The scaling by 100 arises because the original Tränkle & Bailer index is based on just a sample of 100 words).	Higher
Wheeler.Smith	Wheeler and Smith [1954] Readability Measure. $ASL * 10 * (n_{wsy>=2}/n_w)$ where $n_{wsy>=2}$ = the number of words with 2 syllables or more.	Lower
meanSentence Length	Average Sentence Length (ASL). n_w/n_{st}	Lower
meanWord Syllables	Average Word Syllables (AWL). n_{sy}/n_w	Lower

Table 9: All 48 readability measures used in this study - Part 8 of 8.

2.3 Evidence from financial reports readability research in finance indicators

The readability has assumed an increasing importance due to large number of studies using it. Readability establishes a positive relationship with textual comprehension and is precisely why it is relevant and an important tool in accessing financial reports. Some of the instruments that have been used to measure readability are the Fog Index, Bog Index and the word count, mentioned in Section 2.2. Using only these two tools, Li [2008] to establishes a link between readability of 10-K (anual) reports and earnings persistence. Companies with reports with a higher value of Fog Index (lower readability) and with longer text have lower earnings, and companies with more readable reports have more persistent earnings. This suggests that business managers try to hide the company's predictable poor financial results when writing complex documents.

Lawrence [2013] finds that the retail investor's propensity to invest increases in companies with clean and concise reports (measured by the Fog Index). Yu and Miller [2010] show that companies with more readable documents (using the Fog Index as one of their readability measures) have a more dynamic trading activity by small investors than companies with more complex documents. You and Zhang [2009] found, using word count as a measure of readability, that investors react slowly to company reports that are more complex (as measured by word count). On the other hand, De Franco et al. [2015] concluded that the higher the readability of the analyst report, the greater the turnover reaction. Lehavy et al. [2011] in an attempt to relate 10-K report readability to the dispersion and accuracy of financial analysts have come to the conclusion that the number of analysts following the stock increases, the dispersion is larger and there is a smaller accuracy of their earnings forecasts when the readability (measured by the Fog Index) of the text is lower. According to the authors, the inherent costs increase with less understandable 10-Ks reports as more analysts are needed to match investor demand for information about the text. Therefore, financial reports with higher Fog Index values result in an higher dispersion of analysts earnings forecast.

After analyzing this research, it is possible to conclude the perceptible relationship established between readability (measured through several instruments) and the business state of the company, their financial results, future financial perspectives and the decision-making process of the investors.

2.4 TEXT MINING AND NLP

The growing volume of data generated and stored today is largely unstructured or semi-structured, with no clear organization and therefore cannot be used properly to impact a business. In this sense lies the importance of Text Mining, which aims to exploit a great slice of unstructured or semi-structured data: the textual data. As a result of the need to extract standards and knowledge of textual data, the concept of Text Mining has emerged.
The nature of human communication is not trivial for algorithmic development [Gupta et al., 2009]: a word may have different meaning when applied in different contexts or even semantically similar phrases may have words with opposite meanings are some examples.
Thus, when the goal is to apply Text Mining to understand the meaning of textual data, the challenge increases substantially in the case of NLP. NLP is the area responsible for understanding the natural language, extracting its computational meaning and generating natural language. However, [Gupta et al., 2009] points out that, despite these difficulties, computers have the advantage of being able to process large volumes of text, unlike humans, which gives many opportunities to explore and retrieve relevant information.

2.4.1 *Related data analysis areas*

This work is set within the context of several data analysis areas, namely:

1. Data Science is an interdisciplinary field that involves areas such as Data Mining, Data Visualization and Data Analytics. The term covers any set of techniques that aims at extracting data insights, with different purposes: descriptive, predictive, or prescriptive analysis [O'Neil and Schutt, 2013] and understands the whole process from the acquisition of unstructured data to the formation of data products or models [Larson and Chang, 2016].

2. Data Mining is a subprocess in the Data Science pipeline and requires the understanding of the business and data, its preparation, as well as the development and evaluation of the model [Witten et al., 2016]. The purpose of the model is the discovery of new information or relationship between existing information [Sharma et al., 2018] and oftentimes uses Statistics and Machine Learning approaches [Larson and Chang, 2016].

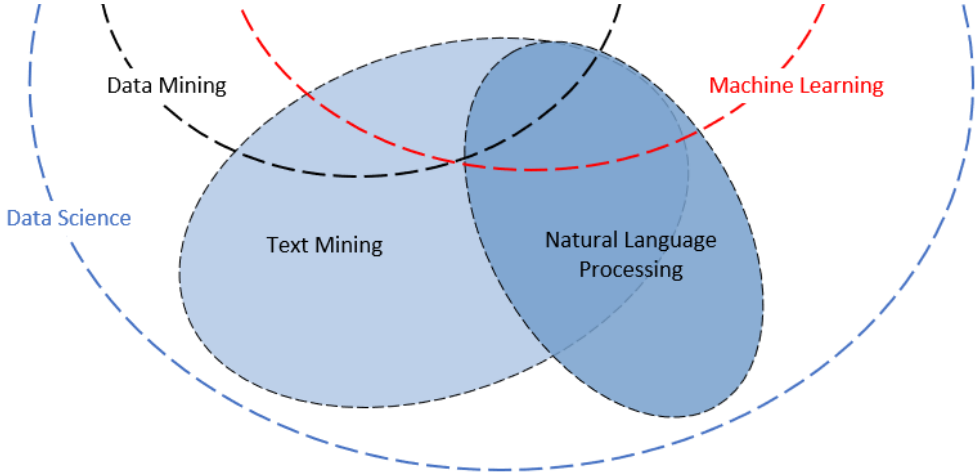


Figure 1: Related areas of Text Mining and NLP.

3. Machine Learning involves the study of algorithms that can extract information automatically (i.e., without direct human guidance). A Machine Learning program learns some task from experience: its performance improves with new data according to some performance measure. Some of these procedures include ideas derived from statistics.

Fig. 1 attempts to explain intuitively the relationship between these areas, showing their interceptions. Check that dashed lines are intended to mean an unclear boundary.

2.4.2 *Topic modelling*

Topic modeling is a type of statistical modeling for discovering similarities in unlabelled texts. Unsupervised Latent Dirichlet Allocation [Blei et al., 2003] is a hierarchical Bayesian model, where topic proportions for a document are drawn from a Dirichlet distribution and words in the document are repeatedly sampled from a topic which itself is drawn from those topic proportions [Zhu et al., 2009].
Usually, studies about topic modelling follow unsupervised approaches, but not all. For example, Ramage et al. [2010] improved the model and made it an supervised approach, learning from training data.
Conditional Random Fields [Nikfarjam et al., 2015] is a method that can take context into account, whereas a discrete classifier predicts a label for a single sample without considering neighboring samples.

2.4.3 *Classification*

Text classification is the Text Mining step that addresses Machine Learning algorithms that aims to discriminate or characterize a piece of text in a particular format value. This value can vary from a number (sentiment analysis), labels (multi-labeling tasks), classes (binary or multi-class tasks). Some examples are: Sentiment analysis, whose goal is to identify the polarity of text content: the type of opinion it expresses. This may take the form of a binary like/dislike rating, or a more granular set of options, such as a star rating from 1 to 5. Some commonly adopted Machine Learning algorithms for text classification are:

1. A Naive Bayes Classifier makes assumptions about how the data (in this case words in documents) is generated and proposes a probabilistic model based on these assumptions. It will then use a set of training examples to estimate the parameters of the model. Bayes rule is used to classify new examples and select the class that most likely has generated the example [Chakrabarti et al., 1997, Allahyari et al., 2017].

2. Nearest Neighbor Classifier is a proximity classifier which uses distance measures to perform the classification. The idea is that documents which belong to the same class are more close to each other based on the similarity measures. The classification of the test document is inferred from the class labels of the similar documents in the training set. If is considered the k-nearest neighbor in the training data set, the approach is called k-nearest neighbor classification and the most common class from these k neighbors is reported as the class label [Han et al., 2001, Allahyari et al., 2017].

3. Decision tree is essentially a hierarchical tree of the training instances, in which a condition on the attribute value is used to divide the data hierarchically. In other words, the decision tree recursively partitions the training data set into smaller subdivisions based on a set of tests defined at each node or branch. Each node of the tree is a test of some attribute, and each branch descending from the node corresponds to one the value of this attribute. An instance is classified by beginning at the root node, testing the attribute by this node and moving down the tree branch corresponding to the value of the attribute in the given instance. This process is then recursively repeated [Allahyari et al., 2017].

4. Support Vector Machines (SVM) are a supervised learning classification algorithms where have been extensively used in text classification problems. SVM are a form of linear classifiers. Linear classifiers in the context of text documents are models that making a classification decision is based on the value of the linear combinations of the documents features [Allahyari et al., 2017].

2.5 FUTURE CRASH RISK

Future crash risk aims to probabilistically measure the risk of a crash. It is essential for assessing the financial health of companies and therefore helps investors make investment decisions.

2.5.1 *What is a crash?*

A crash is a sudden and significant decline in the price of an asset. It is both an economic and a psychological phenomenon, i.e., economically there may be a fall in value perceived by some investors that causes a sudden fall in value that psychologically affects other investors, leading them to follow the trend. Most investors, even if they have no economic reason to sell their shares, are driven to sell for fear that they will lose even more value and are clearly influenced psychologically by the sudden decline. As investors who prefer to sell their shares rather than hold them, investors who could buy also decide not to buy, for the same reasons, thus leading to a growing volume of available stocks with downward demand [Johansen et al., 1999].
For the psychological reasons already mentioned, lack of demand generates more lack of demand: investors will sell their shares and precipitate other investors to sell their shares as well. These investors, who sell their shares after noticing their rapid devaluation, act out of concern that their prices will fall further. Thus, this phenomenon leads to the possibility of a vicious cycle marked by negative crowd behavior.

2.5.2 *Reasons for a higher future crash risk*

The causes of a crash are not deterministic, i.e., there is no set of patterns that clearly identify a crash. However, there are a number of factors where future crash risk is greatest, including:

1. Hide negative information that when released to the market causes extremely negative reactions [Kim et al., 2011].

2. The stock price is inflated [Kim and Zhang, 2016], leading to the creation of a bubble, i.e., the stock price is rising and when the market realizes that it is inflated, the price will go down dramatically, possibly causing a crash.

3. Emerging economies, as emerging equity markets are characterized by excessive volatility and are more likely to have weak corporate governance and may be more susceptible to crashes [Vo, 2019].

4. Companies where institutional investors are most distracted and ignore the attention-grabbing exogenous events [Xiang et al., 2019]. There is an aggravation when companies are state-owned, when CEOs control directors and when there is less coverage by analysts.

Some authors will also identify features in stock returns that can lead to a crash, among them the variance of conditional volatility and the fat tailed distribution of the return series [Bates, 2012]. There is also evidence that some factors can considerably mitigate the risk of a crash, such as factors linked to the reputations of senior management.

2.5.3 *Future crash risk measures*

Since future crash risk is a probabilistic value that measures the likelihood of a crash happening, there are ways to measure it. There are measures created by various authors that allow to estimate future crash risk, for example [Callen and Fang, 2015]:

1. NCSKEW: The Negative Coefficient of Skewness of firm-specific daily returns

2. DUVOL: The Down-to-Up Volatility of firm-specific daily returns.

3. Crash Count: The difference between the number of days with negative extreme firm-specific daily returns and positive extreme firm-specific daily returns.

To calculate these measures, it is necessary to calculate the firm-specific residual daily returns using the expanded market and industry index model for each firm and year [Hutton et al., 2009]:

$$r_{j,t} = \alpha_j + \beta_{1,j} r_{m,t-1} + \beta_{2,j} r_{i,t-1} + \beta_{3,j} r_{m,t} + \beta_{4,j} r_{i,t} + \varepsilon_{j,t} \tag{3}$$

where $r_{i,t}$ is the return on the value-weighted industry index based on 2-digit Standard Industrial Classification SIC codes on day t, $r_{m,t}$ is the return on the CRSP value-weighted market index on day t and rj,t is the return on stock j on day t.
The firm-specific daily return, $R_{j,t}$, is the natural log of (1 plus the residual return from equation 3). Log transforming raw residual returns is used to reduce the positive skew in the return distribution and to help ensure symmetry [Chen et al., 2001]. Thus, the mathematical approach of the measures emerges as:

1. NCSKEW is calculated as the negative of the third moment of each stock's firm-specific daily returns, divided by the cubed standard deviation [Callen and Fang, 2015]. So, for any stock j over the fiscal year T,

$$NCSKEW_{j,T} = \frac{-(n(n-1)^{\frac{3}{2}} \sum R_{j,t}^3)}{((n-1)(n-2)(\sum R_{j,t}^2)^{\frac{3}{2}})} \tag{4}$$

   where n is the number of observations of firm-specific daily returns during the fiscal year T. An increase in the value of NCSKEW corresponds to more likely crashes.

2. DUVOL, is calculated as [Callen and Fang, 2015]:

$$DUVOL_{j,T} = log \left\{ \frac{(n_u - 1) \sum_{DOWN} R_{j,t}^2}{(n_d - 1) \sum_{UP} R_{j,t}^2} \right\} \tag{5}$$

   where n_u and n_d are the number of up and down days over the fiscal year T, respectively. For this measure, it is necessary to separate the days with firm-specific daily returns above and below the period average (e.g., 1 year), shown in the formula as "up" and "down". It is also necessary to calculate the standard deviation for the up and down samples separately and then calculate the logarithmic ratio of the standard deviation of the "down" sample to the standard deviation of the "up" sample. A higher value of DUVOL corresponds to more crash-prone stock.

3. Crash Count, the downside frequencies minus the upside frequencies, uses the number of firm-specific daily returns exceeding standard deviations of 3.09 above and below the firm-specific average daily return over the fiscal year. Stan-

dard deviation 3,09 was used by Hutton et al. [2009] to generate frequencies of 0.1% in the normal distribution. A higher value of Crash Count corresponds to a higher frequency of crashes.

3

DEVELOPMENT

This chapter details all the methodology and development steps to achieve the objectives presented in Section 1.2.

3.1 METHODOLOGY

3.1.1 *The problem and its challenges*

In this project, it is intended to focus on the relationship between the readability measures of financial reports and the correspondent companies' crash risk for each year, which brings several challenges, which are:

1. Understand the financial reports structure and financial information needed to achieve future crash risk.
2. All the data preprocessing, from its unstructured state to its complete structuring. Because the source data is textual, contains formations and a set of anomalies, this step is very laborious and therefore appropriate methodologies should be chosen.
3. Get readability and future crash risk measures.
4. Understanding the relationship between the readability measures of financial reports and the correspondent companies' future crash risk.

3.1.2 *Proposed Approach*

The proposed solution is based on a Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology adapted to this project. CRISP-DM originally has 6 phases [Chapman et al., 2000]:

1. Business Understanding;
2. Data Understanding;
3. Data Preparation;
4. Modeling;
5. Evaluation; and
6. Deployment.

In this project, it is not intended to reach the deployment phase but essentially to evaluate the relationship between the readability measures of financial reports and the correspondent companies' future crash risk. The adapted CRISP-DM methodology is depicted in Fig. 2:

Figure 2: Proposed approach.

3.2 Business understanding

The first step in the adopted CRISP-DM methodology is to understand our goals from a business perspective. It also seeks to discover important factors that may influence the project outcome and to study all the particularities of this project. Thus, it is intended to make a broad study on the project and the surrounding areas.
Since this document is divided into parts, it has been found advantageous to include the knowledge resulting from this phase in the corresponding parts of the document. In particular, the objectives were defined in Section 1.2, the background knowledge was discussed in Chapter 2 and the adopted methodology was explained in Section 3.1.

3.3 Data understanding

3.3.1 *Quality assurance*

Preventive data collection, i.e., avoiding problems with data collection, is the most cost-effective activity to ensure data integrity and data collection [Knatterud et al., 1998]. This process is a set of proactive measures. These measures follow a set of rules that vary from project to project. Thus, the risk of identifying problems and errors early in the process is important and can save many future resources as it will significantly decrease the time spent in case of data collection failures. In this particular project, it was considered very important to select official and reputable sources, that guarantee quality as well as to eliminate failures, such as the failures mentioned by Knatterud et al. [1998]:

1. Uncertainty about time: when this data was created.
2. Uncertainty about methods: how this data was created.
3. Uncertainty about the identity of the persons responsible for the creation and revision of the data.
4. Partial information of some items that should be collected and there is no reason for this lack of information.
5. Vague description of data collection instruments to be used instead of step-by-step instructions on how these data were collected and reviewed.

6. No mechanism identified to document changes in procedures that may evolve throughout the investigation.

It was verified that the sources chosen for this project contain information without this kind of errors and therefore guarantee a smaller amount of errors in the future.

3.3.2 *Importance of ensuring accurate results*

Data accuracy is extremely important to maintaining the integrity of the investigation. It is necessary to obtain datasets from official or reputable organizations, as the sources mentioned in the Section 3.3.3. This data allows to [Most et al., 2003]:

1. Respond appropriately to research questions.
2. Validate the study so that it can be reproduced.
3. True discoveries that may lead other researchers to new discoveries.

The impact of data collection varies from project to project and depends on the nature of the investigation. In this research problem it is necessary to guarantee at least these points, taking into account the characteristics of the financial data and the quality and precision necessary to ensure good results, since these are complex objectives.

3.3.3 *Data collection*

Data collection is one of the most important steps in a Data Science project, as the end results depend greatly on the data collected and the quality and accuracy of the data. Thus, collecting information from relevant sources is essential for finding answers to the research objectives, testing hypotheses to study and evaluating the results. This section aims to explain how the data were obtained. For this project, given the complexity of the objectives, the quality and accuracy of the data to be used will have to be high. Thus, official or reputable sources compatible with all the characteristics mentioned in the previous sections were used, ensuring accurate and high quality data and accuracy. The sources used, as well as the datasets collected from them, are:

1. Securities and Exchange Commission [SEC.gov, 2019], which is an official US agency responsible for enforcing and regulating stock market laws in the United States. This platform was one of the sources used for collecting financial reports.

2. The Software Repository for Accounting and Finance from the University of Notre Dame was used to download the Stage One 10-X Parse Data, which includes clean financial reports from 2006 to 2018 [McDonald, 2019].

3. Another important data source was the Center for Research in Security Prices [CRSP, 2019]. Most datasets used to develop future crash risk metrics were collected from this source, including:

   3.1. Stocks daily;

   3.2. Stocks sample;

   3.3. SIC codes per period;

   3.4. Industry SIC codes; and

   3.5. Industry values.

4. Professor K. French website [Kenneth, 2019] was also a used source. Here, the Fama and French factors were downloaded.

3.3.4 *Describing the data*

In this section, we describe the samples of the dataset.

1. Stocks daily

   For the calculation of future crash risk metrics, several daily stock data are essential, including daily returns (RET). This dataset also contains market capitalization (TCAP) and volume of shares (VOL). The PERMNO variable allows the binding between datasets and the CALDT variable contains the stock date. Thus, this dataset was collected from the CRSP database, containing this information for all United States public companies between 31 December 1925 and 31 December 2018. A sample of this dataset is shown in Table 10.

	PERMNO	CALDT	RET	TCAP	VOL
1	10104	19860312	NA	272023.125	1212967
2	10104	19860313	0.02424242	278617.625	386810
3	10104	19860314	0.03550296	288509.375	178745
4	10104	19860317	-0.02285714	281914.875	87410
5	10104	19860318	-0.02339181	275320.375	99867
6	10104	19860319	-0.01796407	270374.5	70215
7	10104	19860320	-0.006097561	268725.875	55045
8	10104	19860321	-0.03067485	260482.75	50225
9	10104	19860324	-0.006329114	258834.125	64110
10	10104	19860325	0	258834.125	66950

Table 10: Sample of the ”stocks daily” data.

2. Stocks sample

   The daily stocks dataset contains a variable called PERMNO that allows to link to this dataset. This stock sample dataset allows to identify the main characteristics of the companies: the code THICK allows to identify the company for faster search on the SEC platform, CUSIP identifies the North American financial security for the purposes of facilitating clearing and settlement of trades, among other information such as company name. This dataset was collected from the CRSP database. The Central Index Key (CIK) identifies the companies disclosed by the SEC platform and it was obtained manually from the same platform. A sample of this dataset is in Table 11.

	HTICK	PERMNO	CUSIP	CUSIP9	datastream_code
1	ORCL	10104	68389X10	68389X105	719618
2	MSFT	10107	59491810	594918104	719643
3	HON	10145	43851610	438516106	906191
4	KO	11308	19121610	191216100	904282
5	CELG	11552	15102010	151020104	755790
6	XOM	11850	30231G10	30231G102	905039
7	GD	12052	36955010	369550108	907652
8	GE	12060	36960410	369604103	906150
9	CHTR	12308	16119P10	16119P108	68470X
10	GM	12369	37045V10	37045V100	68470T

	co_name	sedol	cusip	cik
1	ORACLE	2661568	68389X105	1341439
2	MICROSOFT	2588173	594918104	789019
3	HONEYWELL INTL.	2020459	438516106	773840
4	COCA COLA	2206657	191216100	21344
5	CELGENE	2182348	151020104	816284
6	EXXON MOBIL	2326618	30231G102	34088
7	GENERAL DYNAMICS	2365161	369550108	40533
8	GENERAL ELECTRIC	2380498	369604103	40545
9	CHARTER COMMS.CL.A	BZ6VT82	16119P108	1091667
10	GENERAL MOTORS	B665KZ5	37045V100	1467858

Table 11: Sample of the ”stocks sample” data.

3. SIC codes per period

   The "SIC codes per period" dataset displays industry codes for each company (linking the company with the PERMNO variable, as explained in the previous points) and for each day because these codes change over time for some companies. Thus, a range of days is presented, with the beginning being the NAMEDT variable and the ending date being the NAMEENDDT variable. The industry code is presented in the SICCD variable. This dataset was collected from the CRSP database. A sample of this dataset is in Table 12.

	PERMNO	NAMEDT	NAMEENDDT	SICCD
1	10104	20040610	20130714	7370
2	10104	20130715	20181231	7372
3	10107	20040610	20181231	7370
4	10145	20040610	20140126	3724
5	10145	20140127	20170129	3714
6	10145	20170130	20181231	5099
7	11308	20040610	20181231	2086
8	11552	20040610	20110126	2890
9	11552	20110127	20181231	2890
10	11850	20040610	20161218	2911

Table 12: Sample of the "SIC codes per period" data.

4. Industry SIC codes

   In order to identify the industry of a company and for each day, it is necessary to convert the SIC code to the industry, which is made from this dataset. Thus, the variable ”min” and ”max” identify the minimum and maximum SIC to be able to belong to one industry and the variable ”number” joins all sub-industries of a larger industry, for example, the first 5 industries belong to a bigger industry. termed ”Agric”. This dataset was collected from the CRSP database. A sample of this dataset is shown in Table 13.

	number	min	max	industry
1	1	100	199	Agric production crops
2	1	200	299	Agric production livestock
3	1	700	799	Agricultural services
4	1	910	919	Commercial fishing
5	1	2048	2048	Prepared feeds for animals
6	2	2000	2009	Food and kindred products
7	2	2010	2019	Meat products
8	2	2020	2029	Dairy products
9	2	2030	2039	Canned preserved fruits vegs
10	2	2040	2046	Flour and other grain mill products

Table 13: Sample of the ”industry SIC codes” data.

5. Industry values

   The SIC for a company and day, obtained in with the "SIC codes per period" dataset, is used to identify the industry (the translation is made using the previous dataset) and thus access the return on the value-weighted industry index, obtained from this dataset. The matrix is accessed by checking the day and the name of the industry. This dataset was collected from the CRSP database. A sample of this dataset is presented in Table 14.

		Agric	Food	Soda	Beer	Smoke	Toys	Fun	Books
1	19260701	0.56	-0.07	-99.99	-1.39	0.00	-1.44	0.62	-1.27
2	19260702	0.29	0.06	-99.99	0.78	0.70	1.46	0.03	0.00
3	19260706	-0.33	0.18	-99.99	-1.74	0.50	-0.96	-0.06	4.27
4	19260707	3.57	-0.15	-99.99	-1.73	-0.12	-0.49	-0.06	-4.10
5	19260708	0.30	1.12	-99.99	-0.15	0.30	-0.49	0.24	0.00
6	19260709	-2.59	0.12	-99.99	-0.07	-0.45	0.00	-0.75	0.85
7	19260710	1.04	1.17	-99.99	0.96	-0.03	0.00	0.92	-0.85
8	19260712	0.28	0.26	-99.99	-1.41	0.38	1.47	0.21	3.42
9	19260713	-1.44	-0.61	-99.99	1.40	-0.21	0.48	0.76	-0.83
10	19260714	0.79	-1.21	-99.99	-0.31	0.21	3.85	-0.47	0.00

Table 14: Sample of the "industry values" data. Showing 8 of 49 industries.

6. Fama and French factors

   The Fama and French factors dataset has market indicators, which will be used to calculate the value-weighted market index, which is essential for calculating future crash risk metrics, as we will see later. This dataset is taken from Professor K. French website [Kenneth, 2019]. A sample of this dataset is in Table 15.

	day	Mkt.RF	SMB	HML	RF
1	19260701	0.1	-0.24	-0.28	0.009
2	19260702	0.45	-0.32	-0.08	0.009
3	19260706	0.17	0.27	-0.35	0.009
4	19260707	0.09	-0.59	0.03	0.009
5	19260708	0.21	-0.36	0.15	0.009
6	19260709	-0.71	0.44	0.56	0.009
7	19260710	0.62	-0.5	-0.15	0.009
8	19260712	0.04	0.03	0.54	0.009
9	19260713	0.48	-0.26	-0.23	0.009
10	19260714	0.04	0.09	-0.48	0.009

Table 15: Sample of the ”Fama and French factors” data.

3.4 DATA PREPARATION

This step is of utmost importance as data preparation defines much of the success of the next steps and is one of the most laborious steps.
The datasets collected in the data collection phase are:

1. Financial reports;
2. Stocks daily;
3. Stocks sample;
4. SIC codes per period;
5. Industry SIC codes;
6. Industry values; and
7. Fama and French factors.

3.4.1 *Select the data*

At the beginning of this phase, all the financial reports of all US listed companies from 2006 to 2018 were available. Since a methodology is being tested, a portion of this data was selected in order to obtain an interesting coherent experimental setup. The annual financial reports (10-K) of the S&P 100 companies were selected during a ten year period, related to the years of 2008 to 2017.
The S&P 100 index is a subset of the S&P 500 index and measures the performance of the largest capitalized companies in the United States. This index includes about 100 of the largest US companies from various industries [Ishares, 2019]. Several investors decide to invest in specific indexes, thus investing in all the companies contained in it, being the S&P 100 one of the most popular US indices. The indexes are quite useful for some investors, depending on their investment strategy, because it allow to mitigate the risk (but also decrease the earnings) since if a company crashes or falls in the value of its shares, remaining index companies can mitigate this loss. However, if one company has a high appreciation the others will mitigate these gains too. In Fig. 3 and Fig. 4, we can see the value of the shares of the S&P 100 index in the last 10 and 30 years, respectively, with monthly plots.

Figure 3: 10 years S&P 100 Stocks Chart (monthly plots), adapted from BarChart [2019].

Figure 4: 30 years S&P 100 Stocks Chart (quarter plots), adapted from BarChart [2019].

There are several types of financial reports, varying in terms of time frame and purpose, including:

1. 10-K reports: These are annual financial information and are mandatory [Investor.gov, 2019].

2. Report 10-Q: Quarter financial information that is not audited [sec, 2019a].

3. 10-KSB Report: Similar to type 10-K, it differs only by targeting small businesses (under $ 75M in stock and $ 50 million in revenue). This type of report has been terminated by the SEC, forcing entities to report in 10-K [sec, 2019b].

Initially, the financial reporting dataset consisted of folders sorted in (1) years, (2) companies and (4) quarters where each quarter's reports were. This setting is shown in Fig. 5. Since the sample of this project uses, as explained, the annual financial reports (10-K) of the S&P 100 companies for 10 years, from 2008 to 2017 being chosen, Python scripts were used to implement the following tasks:

1. The directory structure has been changed, as seen in Fig. 6. 10-K reports could be in any quarter folder.

2. Select only annual reports (10-K), which is the focus of the investigation. This filter was obtained by comparing a section of the name of each report where it indicates the type of report.

3. Select only financial reports from S&P 100 constituent companies. This filter was obtained by obtaining the CIK in the name of each financial report document, subsequently identifying whether the CIK belonged to the S&P 100 index company and thus moving to the new folder.

Thus, these scripts facilitated the automation of the data extraction procedure. For future work, it is possible to look at other years or other companies using the same scripts. Thus, the reproduction of research is greatly facilitated.

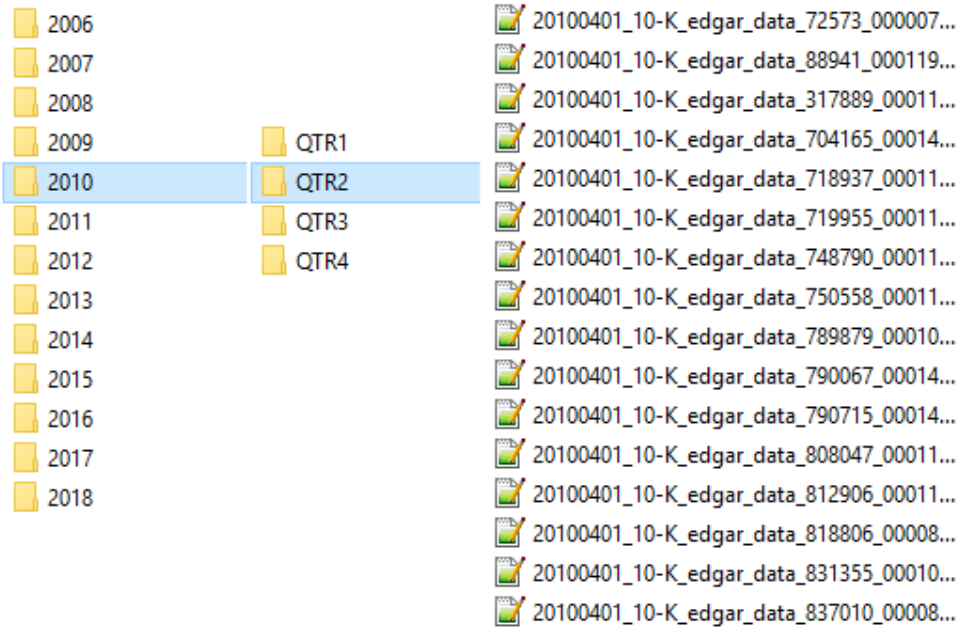


Figure 5: Initial directory of the financial reporting dataset.

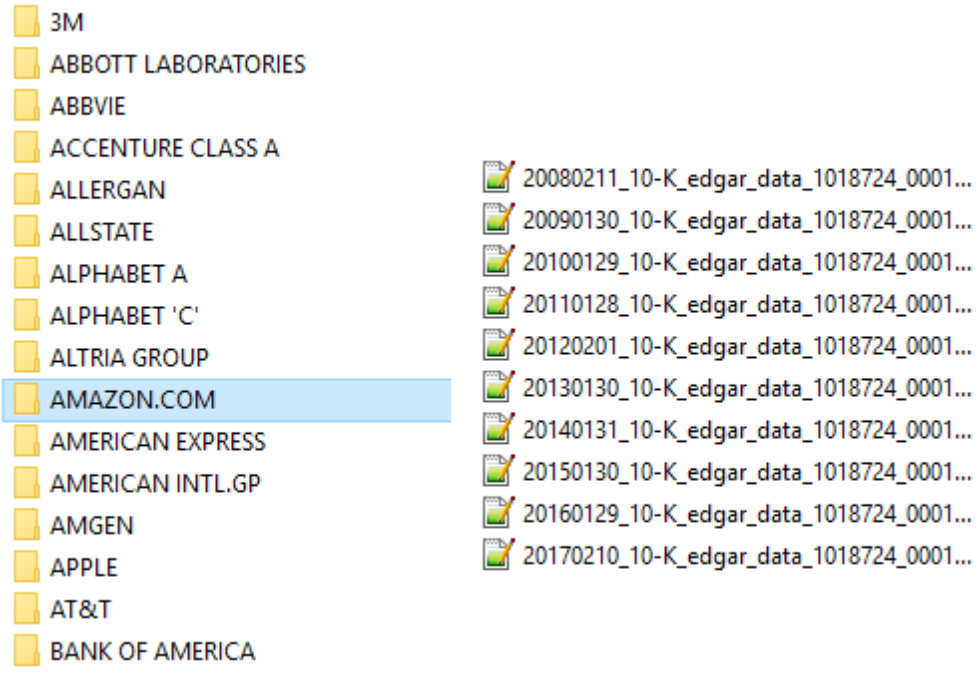


Figure 6: Final directory of the financial reporting dataset.

Data selection by selecting 10 years (2008-2017) of the S&P 100 (total of 102 companies) should yield 1020 (102 * 10) financial reports. Each company should have one annual report per year, totaling 10 financial reports per company. However, there are a total of only 921 financial reports, as seen in Table 16 and Table 17, where it can be seen the number of financial reports by company and by year, respectively. The difference is due to the recent inclusion of some companies in the index or to the stock exchange. This fact was expected and was taken into account.

All the companies:	921
AMAZON.COM	10
ABBOTT LABORATORIES	10
INTERNATIONAL BUS.MCHS.	10
ALLSTATE	10
HONEYWELL INTL.	10
AMGEN	10
AMERICAN EXPRESS	10
AMERICAN INTL.GP.	10
COMCAST A	10
APPLE	10
BERKSHIRE HATHAWAY 'B'	10
VERIZON COMMUNICATIONS	10
BOEING	10
BRISTOL MYERS SQUIBB	10
FEDEX	10
CATERPILLAR	10
CELGENE	10
JP MORGAN CHASE & CO.	10
CISCO SYSTEMS	10
COCA COLA	10
COLGATE-PALM.	10
DANAHER	10
TARGET	10
MORGAN STANLEY	10
WALT DISNEY	10
BANK OF AMERICA	10
CITIGROUP	10
EMERSON ELECTRIC	10
EXXON MOBIL	10
NEXTERA ENERGY	10
GENERAL DYNAMICS	10
GILEAD SCIENCES	10
NVIDIA	10
GENERAL ELECTRIC	10
HALLIBURTON	10
GOLDMAN SACHS GP.	10
HOME DEPOT	10
BIOGEN	10
INTEL	10
JOHNSON & JOHNSON	10
BLACKROCK	10
ELI LILLY	10
UNITED PARCEL SER.'B'	10
LOCKHEED MARTIN	10
LOWE'S COMPANIES	10
MCDONALDS	10
METLIFE	10
CVS HEALTH	10
MICROSOFT	10
3M	10
FORD MOTOR	10
NIKE 'B'	10
WELLS FARGO & CO	10
CAPITAL ONE FINL.	10
OCCIDENTAL PTL.	10
ORACLE	10
EXELON	10
PEPSICO	10
PFIZER	10
CONOCOPHILLIPS	10
ALTRIA GROUP	10
COSTCO WHOLESALE	10
PROCTER & GAMBLE	10
QUALCOMM	10
US BANCORP	10
RAYTHEON 'B'	10
MERCK & COMPANY	10
SCHLUMBERGER	10
SIMON PROPERTY GROUP	10
SOUTHERN	10
AT&T	10
CHEVRON	10
STARBUCKS	10
NETFLIX	10
TEXAS INSTRUMENTS	10
UNION PACIFIC	10
UNITED TECHNOLOGIES	10
UNITEDHEALTH GROUP	10
WALMART	10
MASTERCARD	10
BANK OF NEW YORK MELLON	10
PHILIP MORRIS INTL.	9
VISA 'A'	10
KINDER MORGAN	7
ACCENTURE CLASS A	9
GENERAL MOTORS	8
FACEBOOK CLASS A	5
DUKE ENERGY	10
MONDELEZ INTERNATIONAL CL.A	10
ABBVIE	5
TWENTY-FIRST CENTURY FOX CL.A	0
TWENTY-FIRST CENTURY FOX CL.B	0
BOOKING HOLDINGS	10
MEDTRONIC	3
WALGREENS BOOTS ALLIANCE	3
ALLERGAN	4
DOWDUPONT	0
KRAFT HEINZ	2
ALPHABET A	2
PAYPAL HOLDINGS	2
ALPHABET 'C'	2
CHARTER COMMS.CL.A	10

Table 16: Number of financial reports per company.

	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	Total
Total per year	86	88	89	90	90	92	93	95	99	99	921

Table 17: Number of financial reports per year.

3.4.2 *Financial reports cleaning*

Since the purpose of this project is to analyze the readability of the financial reports text, it is relevant to clean these reports, eliminating all material not necessary for the readability measurement. Much of the content of financial reporting files consists of HTML code, embedded PDF and other artifacts that are of no interest to our purposes [McDonald, 2019]. This content is not only unnecessary and may even prejudice the results. Moreover it also requires a high memory. For example, some records ed initially more than 400MB, but after cleaning the size of the data was reduced to just 5 KB. The steps to clean the financial reports are listed below [McDonald, 2019]:

1. Elimination of document segment <TYPE> tags of ZIP, EXCEL GRAPHIC, JSON and PDF.
2. Elimination of <FONT>, <DIV>, <TR> and <TD> tags.
3. Elimination of all XML code.
4. Elimination of all XBRL segments.
5. Elimination of SEC Header and Footer.
6. \ and \ replacement with a blank space.
7. \&AMP and \&#38 replacement with "&".
8. Elimination of all remaining extended character references.
9. Elimination of all tables. Sometimes some paragraphs are defined with table tags, so each table segment is first stripped of all HTML and then the number of numeric characters against alphabetic characters is compared. Numeric characters / (alphabetic + numeric characters) are calculated and table segments are removed where the result is greater than 10.
10. Translation of exhibits with original tags of "<TYPE> EX-##" to:

    <EX-##>

    ... original text

    </Ex-##>
11. Elimination of all remaining markup tags.

12. Elimination of the excess linefeeds.

13. The HEADER component is eliminated from our analysis.

Clean financial reports were downloaded from the McDonald [2019] source.

3.4.3 *Get readability measures*

At this stage, the aim is to obtain reliable readability measures for financial reports. The cleanup of the financial reports has been completed in Section 3.4.2 and is essential for this phase since no factors outside the meaning of the text are intended to interfere with this assessment.
The following are the steps that were required to obtain readability measures:

1. Read company financial reports for variables, one company at a time, not overloading RAM, by doing the following items and then loading new reports, repeating the process, greatly increasing the speed and decreasing the waste of RAM.

2. Obtain the 48 readability measures for each financial report from the Quanteda library for the R language. These measures are explained in Section 2.2. The name and description of the measures can be seen, for example, in Section 2.2.4.

3. For each row containing the readability measures of a report, also assign the CIK, company name, year and full date, obtaining this information directly from the report name.

4. Save file to rds and csv.

3.4.4 *Get future crash risk measures*

To examine the effect of readability on corporate financial health, the future crash risk indicator discussed in Section 2.5 was used. Thus, several firm-specific future crash risk measures for each year were constructed which are technically explained in Section 2.5.3. This section aims to explain the steps for obtaining these metrics. To this end, it was essential to collect and use several datasets, explained in Section 3.3.3. The algorithmic steps to obtain future crash risk measures include:

1. Read stocks daily, stocks sample, Fama and French factors, SIC codes per period, industry SIC codes and industry values datasets.
2. Read each report individually and apply the following processes.
3. Get variable data binding - PERMNO - from CIK. CIK is found in the name of the financial report document
4. Get the start date and end date of the financial report from its name.
5. Get the daily returns (RET) of the stocks daily dataset from the day after the financial report is published until the day before the next financial report was published.
6. Get $r_{m,t}$ (value-weighted market index): For the same dates of the previous point, get the features Mkt.RF and RF from the Fama and French dataset by applying the formula (Mkt.RF + RF)/100.
7. Get $r_{i,t}$ (return on the value-weighted industry index). To calculate this, obtain the SIC for each of the days mentioned in the previous points. From SIC get the corresponding industry from "industry SIC codes" dataset and then get industry value from "industry values" dataset.
8. Calculate $r_{j,t}$ (return on the value-weighted) using the information presented in Section 2.5.3.
9. Calculate $R_{i,t}$ (firm-specific daily return) as approached in Section 2.5.3.
10. Finally, calculate the future crash risk measures: NCSKEW, DUVOL and Crash Count.

3.5 External analysis with clustering

The collected data was used to relate readability metrics and future crash risk metrics, using the k-means clustering method and subsequent evaluation. The results of this phase can be found in Chapter 4 and this section seeks only to describe the technical aspects that were necessary for the development of this analysis and consequently the production of results, namely:

1. All readability measures that in the Section 2.2.4 are labeled as "More readable when: Lower" have been transformed to the symmetrical, so that all 48 measures are in aligned, where higher values indicate more readable financial reports.

2. Scaling the data previously to the application of the k-means algorithm, saving the scale factors for future unscale inverse operations.

3. The k-means clustering method was used for application in readability measures.

4. Two phases of experiments were done, one with all 48 measures exposed in the Section 2.2.4 and one with only 43 measures, excluding the FOG family, sentence length, or word syllables measures.

5. In each phase, the number of clusters was ranged from 2 to 25 and evaluated using the Gap Statistic, Silhouette, Dunn, Connectivity, APN, AD, ADM and FOM measures to analyze the internal and external dispersion of each cluster, to measure cluster number performance.

6. For each cluster of each experiment, the cluster size and readability measures mean was saved.

7. Externally, for each cluster of each experiment, 3 future crash risk measures (NCSKEW, DUVOL and Crash Count) was calculated, as well as the correlation between them (using all financial reports) and by cluster.

8. The mean and standard deviation of each future crash risk measure for the financial reports of each cluster was calculated as well other cluster features presented in Section 4

9. Correlation between crash risk measures was calculated. Correlation of this measures by cluster was also calculated.

10. Groups of readability measures with equal cluster ranking were selected. For these groups, the average of the selected readability measures was calculated, for each cluster.

11. With the data from the previous points, several graphs have been made which are presented in Section 4.

4

EXPERIMENTS

The proposed solution intends to evaluate the relationship between the readability measures of the financial reports and the future crash risk results of the respectively companies. Readability metrics measure how readable the financial reports are and the future crash risk allows to evaluate the future company's performance. In this project, future crash risk is obtained from the day following the publication of the financial report until the publication of the next report (next year)
In this context, the proposed solution aims to obtain clusters solely from readability measures, having several groups with different readability measures. If readability is related to future crash risk, these groups should also have different future crash risk. Therefore, it is intended to analyze the clusters externally, in the context of future crash risk, looking for differences with some of the 48 readability metrics.
The experiments were performed on index S&P 100 for 10 years (2008-2017), as referenced in the Section 3.4.1. This study focuses on 917 financial reports from 102 different companies as shown in Fig. 16. 48 readability measures (described in Section 2.2.4) and 3 future crash risk measures are evaluated (NCSKEW, DUVOL and Crash Count, described in Section 2.5.3).
With the same data, described in the previous paragraph, 2 case studies were evaluated, having the following differences:

1. Case study 1 in Section 4.1: clustering of all 48 readability measures exposed in the Section 2.2.4 and subsequent external analysis with future crash risk metrics (NCSKEW, DUVOL and Crash Count). For this case study, the data was divided into 6 clusters (as explained in Section 4.1.1).

2. Case study 2 in Section 4.2: clustering of 43 readability measures, excluding the FOG family, sentence length and word syllables measures from the 48 measures of case study 1 and subsequent external analysis with future crash risk metrics

(NCSKEW, DUVOL and Crash Count). For this case study, the data was divided into 65clusters (as explained in Section 4.2.1).

All readability measures that in the Section 2.2.4 are labeled as "More readable when: Lower" have been transformed to the symmetrical, therefore all 48 measures are in tune: higher metrics when financial reports readability is higher. The future crash risk metrics, as explained in Section 2.5.3, is lower for a less crash-prone company. Therefore, it is possible to have an inverse relationship between readability and future crash risk, i.e., a negative correlation, which would mean that there is better financial results (lower future crash risk measures) when the readability is higher (higher readability measures).

4.1 Case study 1: evaluate 6 clusters of 48 readability measures

As explained in the Section 3.5, a clustering method (k-means) was applied to all 48 readability measures exposed in Section 2.2.4 to split the financial reports into groups with similar readability measures. Subsequently, future crash risk is assessed in these groups. This external analysis, in order to analyze the future crash risk of readability clusters, aims to compare the future crash risk measures in the groups to show future crash risk differences between them. All readability measures can be consulted at Section 2.2.4, as well as future crash risk measures in Section 2.5.3.
This section presents the results of the cluster method, starting by evaluating the optimal cluster number, the results of the cluster method (evaluating the division into different readability groups) and then comparing these groups in terms of future crash risk, performing an external analysis.
In color graphics, each cluster is always indicated by the same color, as follows:

1. Cluster 1: red;
2. Cluster 2: orange;
3. Cluster 3: yellow;
4. Cluster 4: fluorescent green (it will be mentioned as green);
5. Cluster 5: cornflower blue (it will be mentioned as blue); and
6. Cluster 6: violet.

4.1.1 *Optimal cluster number evaluation - case study 1*

First, it is intended to evaluate the optimal number of clusters to split the data. This process starts by analysing the optimal number of clusters, testing internal and external validation measures of different number of clusters. Table 18 shows Average Proportion of Non-overlap (APN), Average Distance (AD), Average Distance between Means (ADM), Figure Of Merit (FOM), Connectivity, Dunn and Silhouette validation measures for the number of clusters 2 to the number of clusters 20. In Table 7, a visualization is presented evaluating the ideal number of clusters using the Gap Statistic method. A brief summary of the measures used to evaluate clusters is shown as:

1. Gap Statistic: this metric evaluates the internal dispersion of clusters. It compares the total within intra-cluster variation with their expected values under null reference distribution of the data.

2. APN: this metric analyses the average proportion of observations not placed in the same cluster by clustering based on the full data and clustering based on the data with a single column removed.

3. AD: this metric analyses the average distance between observations placed in the same cluster under both cases (full data set and removal of one column).

4. ADM: this metric analyses the average distance between cluster centers for observations placed in the same cluster under both cases (full data set and removal of one column).

5. FOM: this metric analyses the average intra-cluster variance of the deleted column, where the clustering is based on the remaining (undeleted) columns.

6. Connectivity: this metric corresponds to what extent items are placed in the same cluster as their nearest neighbors in the data space.

7. Dunn index: this metric identifies clusters which are well separated and compacts. It combines dissimilarity between clusters and their diameters to estimate the most reliable number of clusters.

8. Silhouette: this metric analyses how well an observation is clustered and it estimates the average distance between clusters.

Number of Clusters	2	3	4	5	6	7	8	9	10	11
APN	0.00343	0.04473	0.02106	0.11808	0.01158	0.22255	0.11821	0.24765	0.22669	0.355
AD	6.74251	6.08294	5.40579	5.25206	4.70523	4.90695	4.57571	4.65048	4.5473	4.63431
ADM	0.03853	0.36413	0.1629	0.85549	0.0702	1.4531	0.73269	1.41984	1.26158	1.8473
FOM	0.75988	0.65795	0.60346	0.56092	0.51358	0.49993	0.47421	0.45702	0.44877	0.4357
Connectivity	75.5004	155.59802	201.26746	238.39167	276.0873	304.05	365.72302	379.00397	387.69524	396.52063
Dunn	0.0227	0.01886	0.01938	0.02335	0.02287	0.01454	0.01383	0.04771	0.04869	0.04522
Silhouette	0.41246	0.29761	0.29531	0.27694	0.26299	0.25235	0.22901	0.21702	0.21495	0.23039

Number of Clusters	12	13	14	15	16	17	18	19	20
APN	0.27214	0.25579	0.28308	0.36888	0.23365	0.29612	0.36604	0.35373	0.34686
AD	4.38685	4.2886	4.28867	4.3132	3.98407	4.01761	4.03486	3.96618	3.91491
ADM	1.40568	1.32827	1.41405	1.81275	1.11484	1.40322	1.68712	1.59577	1.51159
FOM	0.42417	0.41349	0.41174	0.40283	0.3957	0.38911	0.38034	0.37309	0.37086
Connectivity	438.88968	480.54246	484.32381	443.0746	450.95833	470.40278	509.93175	529.76627	520.74206
Dunn	0.04137	0.04194	0.04194	0.04751	0.05447	0.05447	0.0363	0.03815	0.0283
Silhouette	0.21277	0.20107	0.19965	0.21505	0.2189	0.21753	0.20924	0.20647	0.20351

Table 18: Evaluation of the number of clusters, according to APN, AD), ADM, FOM, Connectivity, Dunn and Silhouette validation measures - case study 1.

To analyze the best number of clusters, an appropriate range of clusters was decided for the methodology. Too few clusters may not be enough to evaluate the relationship, but too many will eventually make each group have few financial reports, making each cluster less significant. Thus, it was decided that a number of clusters between 4 and 10 would be acceptable.

It can be seen from Fig. 7 that the group with the largest Gap Statistic is the group of only one cluster since with one cluster there is only one way to organize the data. Then it can be seen that the measure decreases to the number of clusters 3, increasing thereafter, with a peak in the number of clusters 6, only surpassed from 12 clusters, increasing considerably thereafter. Although Gap Statistic is larger for larger numbers, since clusters are more likely to end up with less dispersion among themselves (this is the relevant factor for this measure), as mentioned in the previous paragraph, a number of clusters between 4 and 10 is acceptable, the number of clusters 6 was chosen, as it corresponds to the largest Gap Statistic among the 4-10 range. It can also be seen from the Table 18 that the number of clusters 6 is one of the optimal numbers for various measures, between the number of clusters 4 to 10. In this sense, the average value of financial reports in each cluster if uniformly distributed is 153 (917 financial reports/6 clusters). However there will be clusters with different sizes.

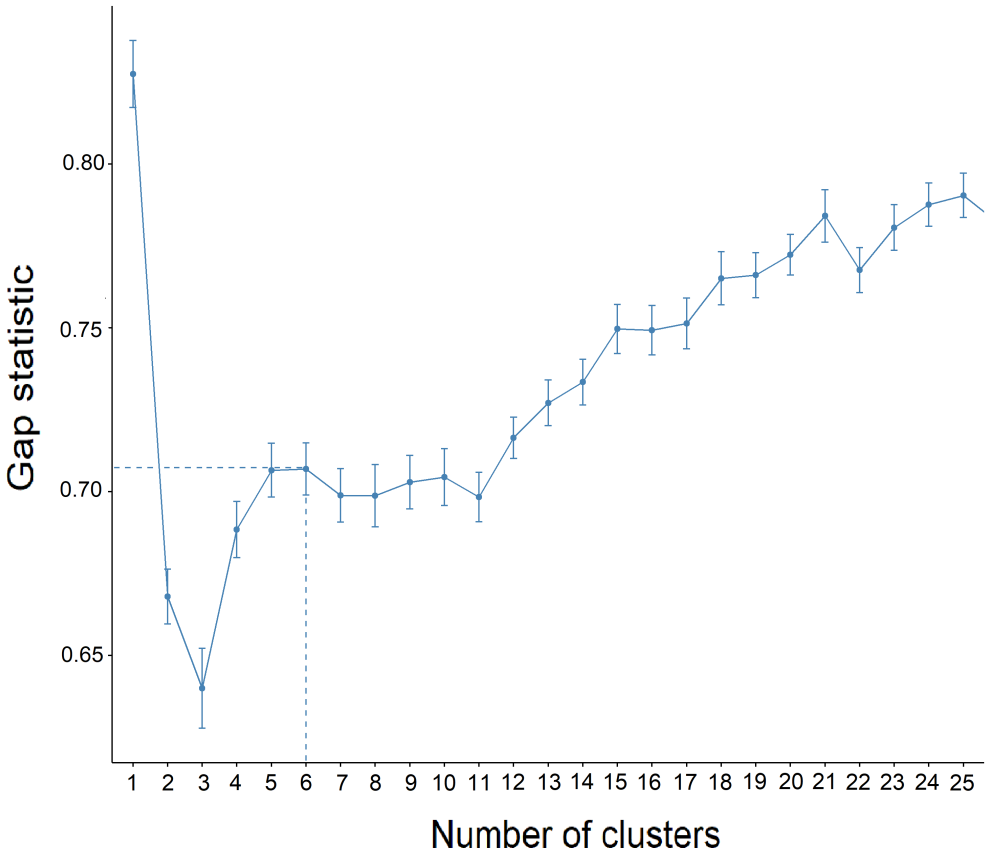


Figure 7: Selection of the number of clusters using the Gap Statistic - case study 1.

4.1.2 *Clustering results - case study 1*

This section presents the results for the 6 readability clusters. For better visualization and discussion, two approaches were made, one focusing on all readability measures, using radar charts (Fig. 9) and another focused on decreasing the complexity of the 48 readability measures, reducing them to only 2 measures, using PCA method (Fig. 8). While radar charts show the results of all 48 readability measures, allowing to discuss and draw accurate conclusions, PCA allows to have a more brief and simpler visualization of the readability clusters.

PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observations from possibly correlated variables (several of the 48 measures are correlated) into a set of linearly uncorrelated variable values called principal components, decreasing the initial number of features [Wold et al., 1987]. Analyzing Fig. 8, it can be seen the result of the PCA method, visualizing the result of the clustering method for readability measures in a simpler way. Each of the two variables (Dim1 and Dim2) represents a certain amount of information contained in the 43 readability measures. The proportion of which is represented by the percentage in the axis labels. Dim1 has 68.6% of the information contained in the initial 48 readability measures and Dim2 contains only 22.1%, noting that Dim1 is responsible for more information. It is also noted that 68.6% + 22.1% = 90.7%, meaning that 9.3% (100% - 90.7%) of the information in this graph has been lost, it is natural since the reduction in the size of the data was large, from 43 features to just 2 features. However, visually it is very worthwhile, since it is clear that at least for the 90.7% of information, the clusters are distinctly different.

Fig. 9 shows the 6 clusters on 4 radar charts (12 readability measures per chart, 48 in total), allowing to analyze the 4 graphs at the same time. Ideally, different clusters should be found, each one corresponding to a particular set of readability measures that, as will be assessed in the next section (where external analysis results are analyzed), would match to a specific future crash risk level (e.g., it would be interesting to find a readability cluster that is often associated with a high degree of future crash risk). Thus, in this section we first want to show and analyze if the clustering method was effective. Taking this into consideration, it appears that the 6 clusters have obvious differences and there is no cluster that has a similar pattern to others for the 43 readability measures analysed in Fig. 9.

The clusters analyzed in Fig. 9 show a differentiation of them, considering the 43 readability measures visualized in the 4 radar charts of this figure. However, due to

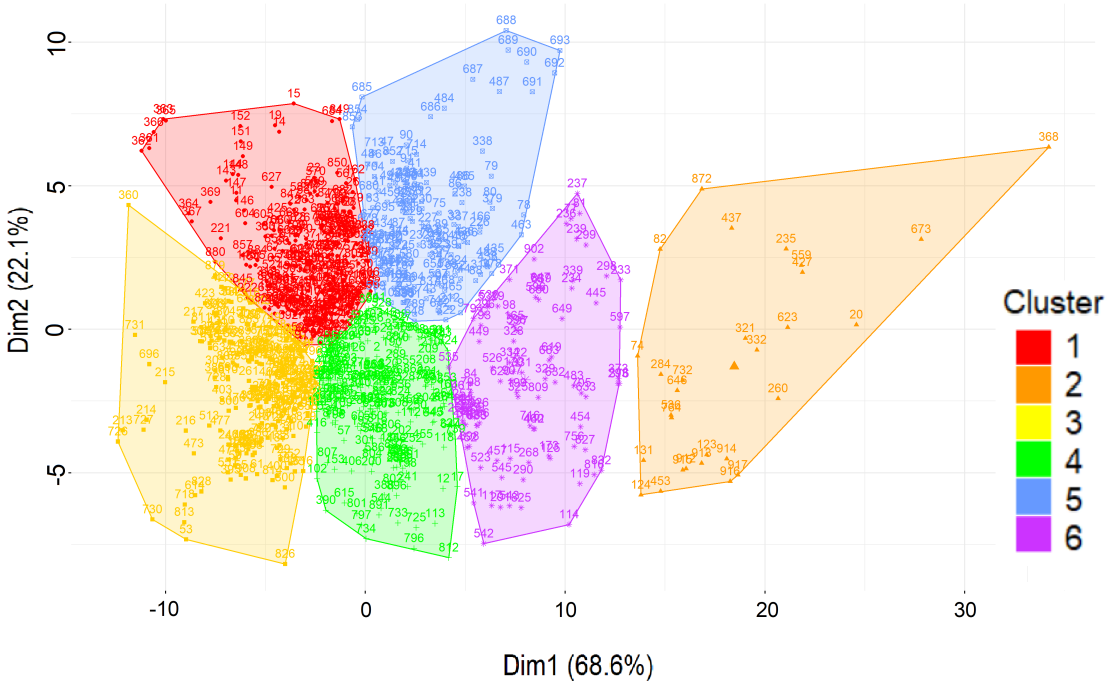


Figure 8: PCA results. The points inside the clusters represent the financial reports - case study 1 (with 6 clusters).

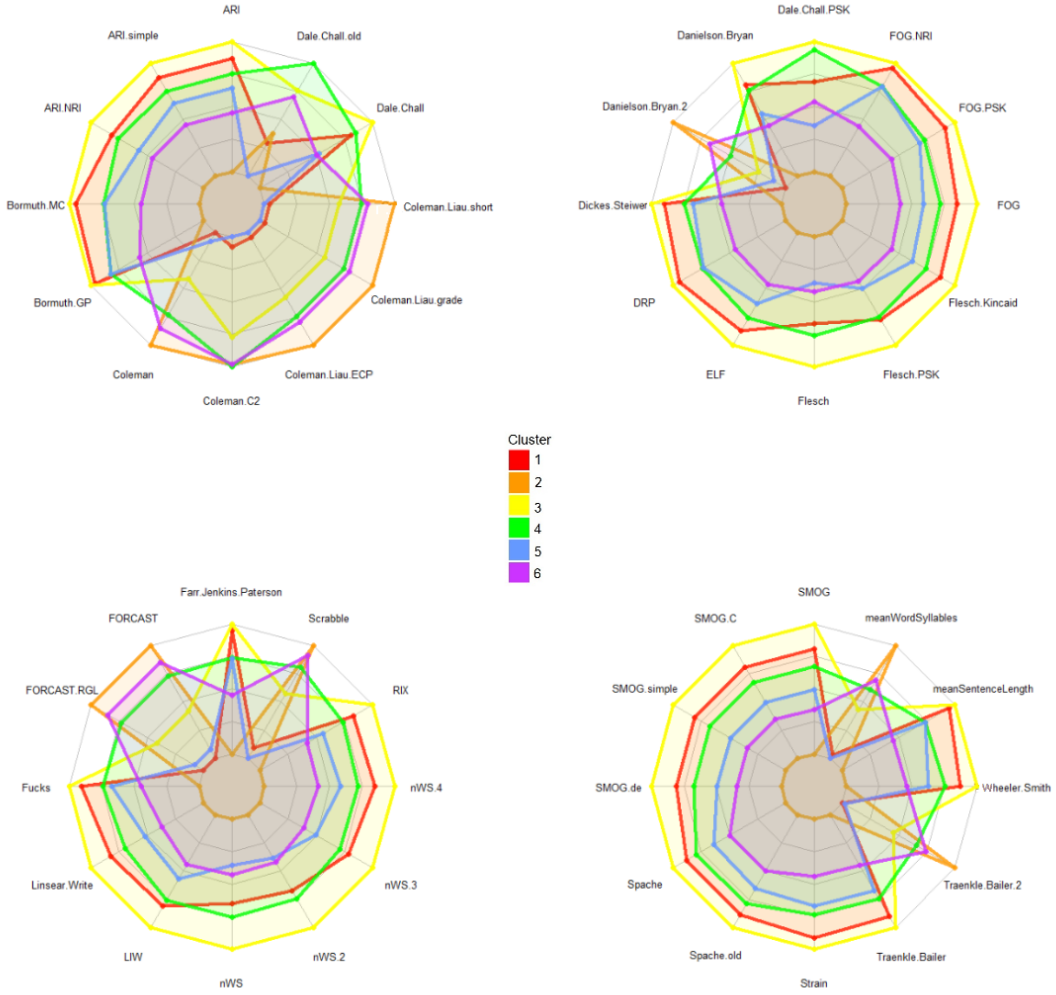


Figure 9: Radar charts - case study 1 (with 6 clusters). The closer to the outer edge of the chart a measure is, the more readable the cluster is.

the number of features, the view in Fig. 8 becomes more favorable, clearly seeing totally distinct clusters. In this sense, the PCA was of great importance, reducing the initial spectrum from 48 characteristics to only 2, Dim1, containing 68.6% of the initial information of the 43 readability measures and Dim2, containing 22.1% of this information, meaning that Dim1 represents most of the information obtained in this graph. This view proves the success of the k-means clustering method by splitting the clusters properly.

4.1.3 *External analysis results - case study 1*

After observing and analyzing the clusters for the 48 readability measures in Section 4.1.2, it was concluded that it created suitable groups. Taking this into consideration, it is now possible to analyze the clusters in terms of future crash risk, evaluating future crash risk measures in each cluster, externally. Readability metrics measure how readable the financial reports are and the future crash risk allows to evaluate the future company's performance. In this project, future crash risk is obtained from the day following the publication of the financial report until the publication of the next report (next year). Thus, this phase goes through characterizing the readability clusters with future crash risk and realizing what distinguishes them.
As explained, the financial characteristic analyzed in this project is the future crash risk. The financial graphs in Fig. 10, Fig. 11 and Fig. 12 show the future crash risk measures of the 6 clusters analyzing 2 future crash risk measures per graph. The size of the circles is proportional to the size of the clusters. Therefore, it is possible to have the notion of the size of each cluster from these views, also using this feature for their comparison. It is also possible to visualize the correlation between the different future crash risk measures in Table 25 and by cluster in Table 20.
The following points list commonalities between readability and future crash risk measures, taking into account the PCA method.

1. Focusing on the readability clusters depicted in Fig. 8, representing the clusters obtained from readability measures, it can be seen that Cluster 2 is isolated from others, in Dim1. As already explained (Section 4.1.2), the method PCA, represented in this figure (Fig. 8), contains 2 features (resulting from the initial 48 readability measures), Dim1 and Dim2, as depicted in the image. Dim1 contains 68.6% of the initial information of the 48 readability measures and Dim2 contains 22.1% of the information, meaning that Dim1 represents most of the information

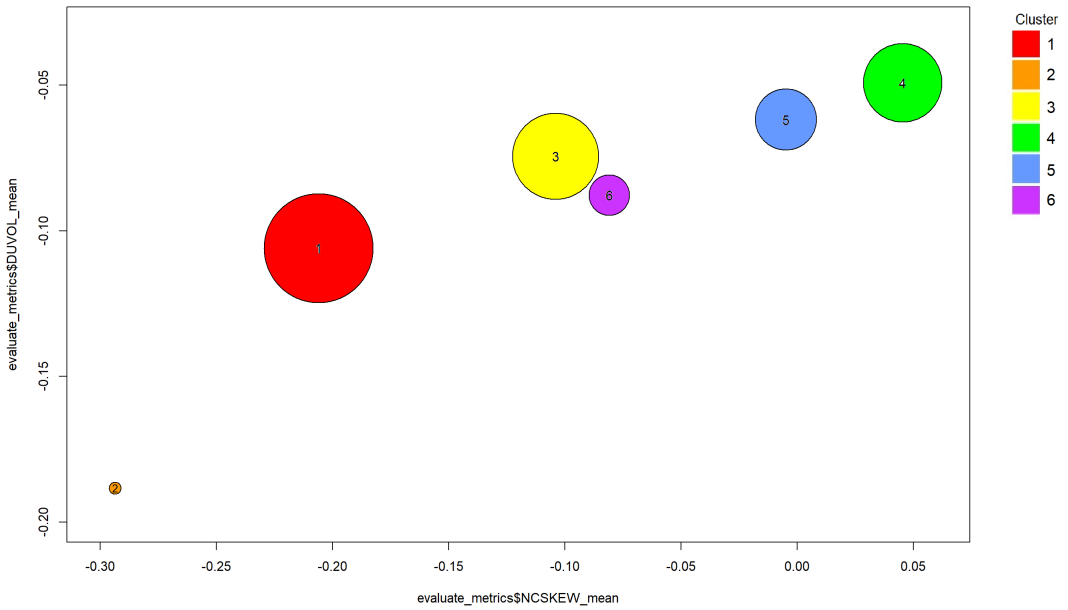


Figure 10: NCSKEW and DUVOL per cluster - case study 1 (with 6 clusters). Circles are proportional to the size of the cluster.

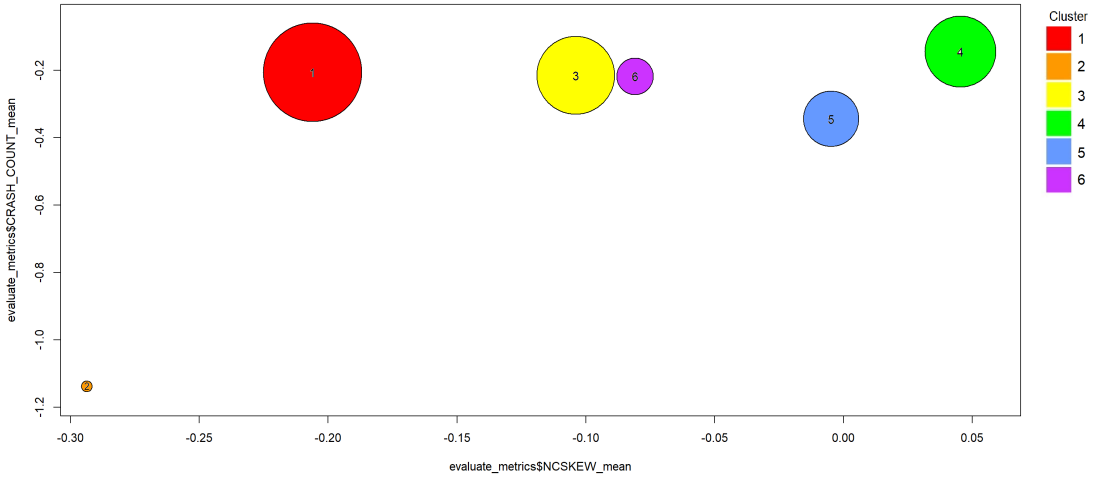


Figure 11: NCSKEW and Crash Count per cluster - case study 1 (with 6 clusters). Circles are proportional to the size of the cluster.

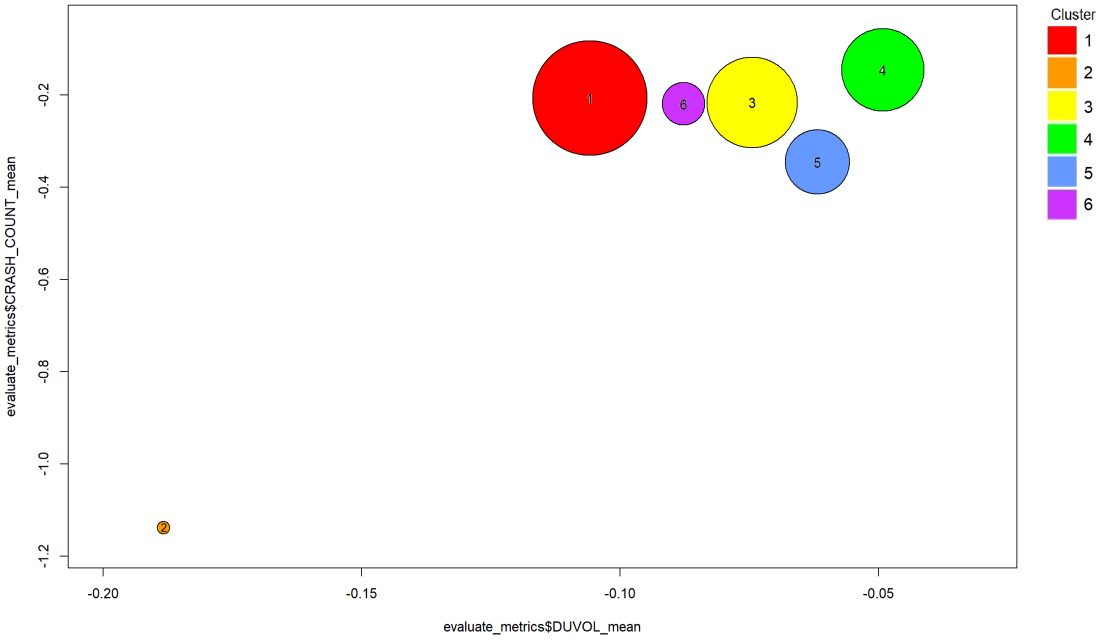


Figure 12: DUVOL and Crash Count per cluster - case study 1 (with 6 clusters). Circles are proportional to the size of the cluster.

obtained in this graph. Therefore, this cluster is isolated in the most of the information, Dim1. Analysing the 3 future crash risk metrics (NCSKEW, DUVOL and Crash Count), in Fig. 10, Fig. 11 and Fig. 12, Cluster 2 always appears further away from all other clusters. Cluster 2 has the lowest future crash risk mean (-0.54, as seen in Table 19) and the lowest NCSKEW, DUVOL and Crash Count. This is the cluster that identifies companies with the lower future crash risk. As mentioned, this cluster also has a differentiation in most readability information (68.6 %), as seen in Fig.8, therefore this may indicate a relationship between some readability measures (68.6 %) and future crash risk measures, but it does not show what type of relationship is (how do the measures affect each other).

2. It can also be seen from Fig. 8 that clusters 1 and 3, represented by red and yellow, respectively, are more or less aligned in the Dim1. As explained, Dim1 contains more information than Dim2 and therefore it is more representative: Dim1 contains 68.6% of the information and Dim2 contains 22.1% of the information. These clusters (1 and 3) are similar in Dim1 (most of the information).

	size_clusters	NCSKEW_mean	DUVOL_mean	CRASH_COUNT_mean	mean_all_CR_metrics
Cluster *1*	257	-0.20609	-0.10592	-0.20623	-0.17274
Cluster *2*	29	-0.29361	-0.18837	-1.13793	-0.53997
Cluster *3*	204	-0.10392	-0.07446	-0.21569	-0.13135
Cluster *4*	186	0.04535	-0.04923	-0.14516	-0.04968
Cluster *5*	145	-0.00475	-0.06186	-0.34483	-0.13714
Cluster *6*	96	-0.08089	-0.08776	-0.21875	-0.12913

Table 19: Cluster size and external features mean - case study 1 (with 6 clusters).

	NCSKEW.and.DUVOL	NCSKEW.and.CRASH_COUNT	DUVOL.and.CRASH_COUNT
Cluster 1	0.9065	0.56178	0.69846
Cluster 2	0.85952	0.47521	0.59232
Cluster 3	0.88186	0.51807	0.67973
Cluster 4	0.90824	0.56015	0.68522
Cluster 5	0.93033	0.64916	0.74282
Cluster 6	0.88098	0.48087	0.59489

Table 20: Correlation of future crash risk measures per cluster - case study 1 (with 6 clusters).

Analysing future crash risk, in Table 19 and specifically the Crash Count metric, these clusters have a very close value too (both have -0.2 and in this metric all clusters values vary between -0.1 and -1.1).

Other relationships can be seen by analyzing the cluster rankings for readability measures (these can be grouped into similar groups) and then analyzing the cluster rankings for future crash risk measures. Comparing both rankings is important, since a similar clustering order between several readability measures and some future crash risk measures can be a great indicator of the relationship between these measures. Thus, since 48 readability measures are analysed, it is intended to group the measures in similar groups and considering their cluster order. After this, this groups will be compared with the results of future crash risk rankings by measure. These groups can be compared in terms of readability and future crash risk at the same time for some clusters, as it will be seen. Then, it is presented the readability groups and analyzed the points that these groups have in common considering their calculation formula, using the information from the Table 2 to the Table 9:

1. Group 1 (measures sorted in descending order in Table 21): 33 metrics including ARI family (3 metrics), Bormuth family (2 metrics), Dale.Chall, Dale.Chall.old,

Danielson.Bryan, Dickes.Steiwer, DRP, ELF, Flesch.PSK, Flesch.Kincaid, FOG family (3 metrics), Farr.Jenkins.Paterson, Fucks, Linsear.Write, LIW, nWS.3, nWS.4, RIX, SMOG family (4 metrics), Spache family (2 metrics), Strain, Traenkle.Bailer, Wheeler.Smith and meanSentenceLength.

This group contains major emphasis on Average Sentence Length (ASL) and Average Word Length (AWL), also considering the size of syllables of a sentence, n_{wsy}, (e.g., ELF, Farr.Jenkins.Paterson, Flesch, Flesch.PSK, FOG), with different variations (e.g., word count with 1, 2 or 3 syllables per word, depending on the metric). Less frequently, another components are also used as: number of "difficult" words not matching the Dale-Chall list of "familiar" words, n_{wd} (e.g., Dale.Chall.PSK), number of words, n_w (e.g., Dale.Chall. PSK, Flesch, Flesch.PSK), number of characters, n_c (e.g., Danielson.Bryan), the number of blanks, n_{blank} (e.g., Danielson.Bryan) and number of sentences, n_{st} (e.g., Danielson.Bryan, ELF, FOG). See notation in Section 2.2.4.

2. Group 2 (measures sorted in descending order in Table 22): 4 metrics including Dale.Chall.PSK, Flesch, nWS and nWS.2.

   The components of this group are similar to that of Group 1, but also using the number of characters per word, n_{wchar}, (e.g., nWS and nWS.2 using the number of words with 6 characters or more, $n_{wchar>=6}$) and without using Average Word Length (AWL), in contrast to Group 1 which has a great use of this component.

3. Group 3 (measures sorted in descending order in Table 23): 11 metrics including Coleman family (5 metrics), Danielson.Bryan.2, FORCAST family (2 metrics), Scrabble, Traenkle.Bailer.2 and meanWordSyllables.

   This group does not use Average Sentence Length (ASL), unlike previous groups. This group is based on n_{wsy} (e.g., Coleman, Coleman.C2, FORCAST), n_w (e.g., Traenkle.Bailer2, Coleman.Liau.ECP, FORCAST), n_{st} (e.g., Danielson.Bryan2). Other less commonly used components are the number of prepositions, n_{conj} (e.g., Traenkle.Bailer2) and the number of conjunctions, n_{nconj} (e.g., Traenkle.Bailer2). See notation in Section 2.2.4.

Looking at Table 19, it is now possible to ranking the clusters regarding the future crash risk metrics. The three future crash risk metrics have some consistency in the order they define clusters, since they are partially correlated, as shown in Fig. 20. However, they all have a slightly different order. The cluster rankings of the future crash risk measures and their mean are shown:

	mean_all_scale	ARI	ARI.simple	ARI.NRI	Bormuth.MC	Bormuth.GP	Coleman.C2
Cluster 2	0.903818981212074	-16.7512242902	-74.1407056435331	-14.9519561823674	-4.90791079514032	-317833396.599133	29.7727108403473
Cluster 1	0.396956189885797	-17.6268723502758	-75.8441373557156	-15.9068397699208	-5.04793540703877	-345859839.311022	27.659193533484
Cluster 5	-0.424284332365531	-19.0456343554397	-78.768574466159	-16.5799683277227	-5.95769808030393	-566879127.938654	30.8677624725863
Cluster 4	-0.647859029205778	-19.4318784296159	-79.4627517348865	-17.3051299759275	-5.78145744186998	-518070882.83064	27.5212798296548
Cluster 3	-2.49353639380871	-22.6204411560521	-85.9337184601294	-19.3572808239076	-7.39568754954055	-1092843021.83268	30.2093970628109

	Coleman.Liau.ECP	Coleman.Liau.grade	Coleman.Liau.short	Dale.Chall	Dale.Chall.old	Dale.Chall.PSK	Danielson.Bryan
Cluster 2	32.4116732469947	-14.1830218836305	-14.1841290246686	5.05500988805453	-11.7069402865609	-9.78922327318302	-7.59805327345514
Cluster 1	29.8655002329838	-14.8806834741615	-14.8818193553696	3.72173741054028	-11.8844316395526	-9.93485856233233	-7.84538150253667
Cluster 5	33.7056934753611	-13.8284551649772	-13.8296311131297	3.55350995354828	-11.6008387263788	-9.83904115284935	-8.00484462251674
Cluster 4	29.8122741249179	-14.895267640676	-14.896458527665	1.62744359826627	-11.9850094661213	-10.0971090020247	-8.21092430171681
Cluster 3	33.8682680277124	-13.7839090873347	-13.7851538735821	-0.687491806311862	-11.8177325850341	-10.1725367118464	-8.69251744162087

	Dickes.Steiwer	DRP	ELF	Farr.Jenkins.Paterson	Flesch	Flesch.PSK	Flesch.Kincaid
Cluster 2	-487.268501906187	-590.791079514032	-12.314018118133	-57.5905966680324	25.7300595054872	-8.15320366268472	-16.2617711293443
Cluster 1	-504.083486520699	-604.793540703878	-13.0885839470271	-58.3057612512232	21.9556696861542	-8.37193126302863	-16.9565986028307
Cluster 5	-547.906651247474	-695.769808030393	-14.2848373075211	-63.1088520869358	22.4481296194493	-8.45655219150523	-18.077100066071
Cluster 4	-549.595161374439	-678.145744186998	-14.7807862932817	-62.1583448425356	18.1535045745845	-8.66469052309355	-18.4316825225759
Cluster 3	-638.858028579711	-839.568754954056	-17.5609714918325	-70.7164144982077	15.4275210065973	-9.00817093178139	-20.9190308839037

	Fucks	Linsear.Write	LIW	nWS	nWS.2	nWS.3	nWS.4
Cluster 2	-140.28077175184	-25.727621272035	-60.7601169179304	-11.9092778231108	-12.274341357734	-11.1307374585646	-12.0132015453439
Cluster 1	-146.94127349088	-27.8249109637884	-63.1951456006943	-12.5757936557005	-12.9094584503125	-11.6563381753554	-12.5603975464401
Cluster 5	-166.218293164138	-30.4299736146278	-64.8203570116465	-12.4698993062539	-12.8830263292781	-12.0332759402996	-13.3363682260649
Cluster 4	-166.879409784485	-31.9316956137474	-66.6136595593328	-13.1697830991271	-13.5201454250369	-12.4302218875276	-13.616229757816
Cluster 3	-203.641651771504	-36.9783519160731	-71.8940950073065	-13.598135782398	-14.018862732702	-13.3690902987795	-15.2421177258849

	Scrabble	SMOG	SMOG.C	SMOG.simple	SMOG.de	Spache	Spache.old
Cluster 2	1.61695883825924	-17.6113502101116	-16.9240382424727	-16.8851871621396	-11.8851871621396	-8.28464549357756	-9.21066367935265
Cluster 1	1.60708097454722	-18.191143168049	-17.4723267815942	-17.4410768629425	-12.4410768629425	-8.46144718336637	-9.40564468688926
Cluster 5	1.62229641297061	-18.8812357438216	-18.125572872742	-18.1027188339613	-13.1027188339613	-8.70114585074658	-9.72451903256985
Cluster 4	1.60657344204716	-19.2650602364916	-18.4891889388807	-18.4707193063199	-13.4707193063199	-8.85309925743331	-9.87001402692001
Cluster 3	1.62439854731758	-20.4874966153601	-19.6484724420417	-19.6427580204794	-14.6427580204794	-9.50407826326308	-10.672321519741

	Strain	Traenkle.Bailer	Wheeler.Smith
Cluster 2	-14.4994649230366	-545.210476167318	-123.14018118133
Cluster 1	-15.1651917028954	-562.111170685785	-130.885839470271
Cluster 5	-17.2294588947057	-605.710411527726	-142.848373075211
Cluster 4	-17.2803638238101	-607.457159684075	-147.807862932817
Cluster 3	-21.1735715276141	-695.147126803352	-175.609714918325

Table 21: Readability measures Group 1 for case study 1 (with 6 clusters)

	mean_all_scale	Coleman	Coleman.C2	Coleman.Liau.ECP	Coleman.Liau.grade	Coleman.Liau.short
Cluster 3	1.19030347447876	33.149111088197	30.2093970628109	33.8682680277124	-13.7839090873347	-13.7851538735821
Cluster 5	1.09487710316762	32.9196068716177	30.8677624725863	33.7056934753611	-13.8284551649772	-13.8296311131297
Cluster 2	0.144214640771086	30.6332258569585	29.7727108403473	32.4116732469947	-14.1830218836305	-14.1841290246686
Cluster 4	-0.520766925434813	29.0339187545841	27.5212798296548	29.8122741249179	-14.895267640676	-14.896458527665
Cluster 1	-0.769273282132506	28.4362510876149	27.659193533484	29.8655002329838	-14.8806834741615	-14.8818193553696

	Danielson.Bryan.2	FORCAST	FORCAST.RGL	Scrabble	Traenkle.Bailer.2
Cluster 3	81.5319540718303	-11.6745219664887	-11.2719741631376	1.62439854731758	-307.223023133906
Cluster 5	80.4225375360597	-11.7012085033003	-11.3013293536303	1.62229641297061	-309.535666241896
Cluster 2	78.9553839572913	-11.9670667608188	-11.5937734369007	1.61695883825924	-316.538257838831
Cluster 4	78.502823030841	-12.1530327029553	-11.7983359732509	1.60657344204716	-325.784244485177
Cluster 1	77.9034527375872	-12.2225289433006	-11.8747818376306	1.60708097454722	-326.896623996348

Table 22: Readability measures Group 2 for case study 1 (with 6 clusters)

	mean_all_scale	Bormuth.MC	Bormuth.GP	DRP	Dickes.Steiwer	Farr.Jenkins.Paterson
Cluster 2	0.73456479879323	-4.90791079514032	-317833396.599133	-590.791079514032	-487.268501906187	-57.5905966680324
Cluster 1	0.550958529793262	-5.04793540703877	-345859839.311022	-604.793540703878	-504.083486520699	-58.3057612512232
Cluster 4	-0.410866982298068	-5.78145744186998	-518070882.83064	-678.145744186998	-549.595161374439	-62.1583448425356
Cluster 5	-0.641961256585344	-5.95769808030393	-566879127.938654	-695.769808030393	-547.906651247474	-63.1088520869358
Cluster 3	-2.52751464415159	-7.39568754954055	-1092843021.83268	-839.568754954056	-638.858028579711	-70.7164144982077

Table 23: Readability measures Group 3 for case study 1 (with 6 clusters)

1. Sorting clusters by NCSKEW mean, ascending form: [2,1,3,6,4,5]. Color order: [orange, red, yellow, violet, green, blue].

2. Sorting clusters by DUVOL mean, ascending form: [2,1,6,5,4,3]. Color order: [orange, red, violet, blue, green, yellow].

3. Sorting clusters by Crash Count mean, ascending form: [2,5,6,3,1,4]. Color order: [orange, blue, violet, yellow, red, green].

To show the evidence of the relationship of these readability measures groups with future crash risk measures, it will be analyzed by the order of clusters for the future crash risk measures mentioned above and their relationship with the proposed groups. As already listed, the ascending order of clusters with respect to future crash risk measures is again mentioned: (NCSKEW: [2,1,3,6,4,5]), (DUVOL: [2,1,6,5,4,3]) and (Crash Count: [2,5,6,3,1,4]). These sortings will be used as follows, analysing the relationship of readability measures groups and future crash risk measures:

1. Group 3 with NCSKEW: Considering the clusters order [2,6,4,5] (without cluster 1 and 3), painted [orange, violet, green and blue], these clusters follow this order for readability metrics Group 3 (in descendent order) and the future crash risk metric NCSKEW (in ascending order).

2. Group 3 with DUVOL: Considering the clusters order [2,6,4,3] (without cluster 1 and 5), painted [orange, violet, green, yellow], these clusters follow this order for readability metrics Group 3 (in descendent order) and the future crash risk metric DUVOL (in ascending order).

3. Group 3 with Crash Count: Considering the clusters order [2,6,3,1] (without cluster 4 and 5), painted [orange, violet, yellow and red], these clusters follow this order for readability metrics Group 3 (in descendent order) and the future crash risk metric Crash Count (in ascending order).

4. Group 2 with NCSKEW: Considering the clusters order [1,6,5] (without cluster 2, 3 and 4), painted [red, violet, blue], these clusters follow this order for readability metrics Group 2 (in descendent order) and the future crash risk metric NCSKEW (in ascending order).

5. Group 2 with DUVOL: Considering the clusters order [1,6,5] (without cluster 2, 3 and 4), painted [red, violet, blue], these clusters follow this order for readability

metrics Group 2 (in descendent order) and the future crash risk metric DUVOL (in ascending order).

6. Group 1 with Crash Count: Considering the clusters order [3,1,4] (without cluster 2, 5 and 6), painted [yellow, red, green], these clusters follow this order for readability metrics Group 1 (in descendent order) and the future crash risk metric Crash Count (in ascending order).

4.2 CASE STUDY 2: EVALUATE 5 CLUSTERS OF 43 READABILITY MEASURES

In this section, the second experiment will be done, evaluating the relationship between 43 readability measures (excluding the FOG family, sentence length and word syllables measures from the 48 measures of case study 1) and future crash risk (NCSKEW, DUVOL and Crash Count). As explained in the Section 3.5, a clustering method (k-means) was applied to readability measures to split the financial reports into groups with similar readability measures. Subsequently, future crash risk is assessed in these groups. This external analysis, in order to analyze the future crash risk of readability clusters, aims to compare the future crash risk measures in the groups to show future crash risk differences between them.
While in the first experiment all readability measures were used, in this experiment some common measures were removed to analyze the reaction of the clustering method and the results of external analysis on a different set of measures. Thus, the 3 FOG family measures (FOG, FOG.PSK, FOG.NRI) were removed, as well as the sentence length and word syllables measures. All readability measures can be consulted at Section 2.2.4, as well as future crash risk measures in Section 2.5.3.
This section presents the results of the cluster method, starting by evaluating the optimal cluster number, the results of the cluster method (evaluating the division into different readability groups) and then comparing these groups in terms of future crash risk, performing an external analysis.
In color graphics, each cluster is always indicated by the same color, as follows:

1. Cluster 1: red;
2. Cluster 2: orange;
3. Cluster 3: yellow;
4. Cluster 4: fluorescent green (it will be mentioned as green); and
5. Cluster 5: cornflower blue (it will be mentioned as blue).

4.2.1 *Optimal cluster number evaluation - case study 2*

First, it is intended to evaluate the optimal number of clusters to split the data. This process starts by analysing the optimal number of clusters, testing internal and external validation measures of different number of clusters. Table 24 shows Average

Proportion of Non-overlap (APN), Average Distance (AD), Average Distance between Means (ADM), Figure Of Merit (FOM), Connectivity, Dunn and Silhouette validation measures for the number of clusters 2 to the number of clusters 20. In Fig. 13, a visualization is presented evaluating the ideal number of clusters using the Gap Statistic method. A brief summary of the measures used to evaluate clusters is shown as:

1. Gap Statistic: this metric evaluates the internal dispersion of clusters. It compares the total within intra-cluster variation with their expected values under null reference distribution of the data.

2. APN: this metric analyses the average proportion of observations not placed in the same cluster by clustering based on the full data and clustering based on the data with a single column removed.

3. AD: this metric analyses the average distance between observations placed in the same cluster under both cases (full data set and removal of one column).

4. ADM: this metric analyses the average distance between cluster centers for observations placed in the same cluster under both cases (full data set and removal of one column).

5. FOM: this metric analyses the average intra-cluster variance of the deleted column, where the clustering is based on the remaining (undeleted) columns.

6. Connectivity: this metric corresponds to what extent items are placed in the same cluster as their nearest neighbors in the data space.

7. Dunn index: this metric identifies clusters which are well separated and compacts. It combines dissimilarity between clusters and their diameters to estimate the most reliable number of clusters.

8. Silhouette: this metric analyses how well an observation is clustered and it estimates the average distance between clusters.

To analyze the best number of clusters, an appropriate range of clusters was decided for the methodology. Too few clusters may not be enough to evaluate the methodology properly, but too many will eventually make each group have few financial reports, making each cluster less significant. Thus, it was decided that a number of clusters between 4 and 10 would be acceptable.

Number of Clusters	2	3	4	5	6	7	8	9	10	11
APN	0.01241	0.01335	0.02155	0.1236	0.0223	0.2105	0.26564	0.38499	0.36416	0.28919
AD	6.40033	5.67766	5.09704	4.93593	4.46158	4.57076	4.50552	4.57706	4.47603	4.27406
ADM	0.13093	0.11298	0.16681	0.90891	0.13998	1.2144	1.39537	2.02303	1.82758	1.44578
FOM	0.76372	0.65501	0.60299	0.56964	0.51461	0.49356	0.47246	0.45478	0.44863	0.4406
Connectivity	89.07976	179.24841	210.45714	239.99048	277.69484	296.8377	373.85992	336.13214	349.05714	388.12778
Dunn	0.01579	0.01264	0.01905	0.0285	0.02444	0.02388	0.02389	0.03323	0.04865	0.02628
Silhouette	0.40619	0.28019	0.29622	0.27764	0.26984	0.25653	0.22512	0.24054	0.2431	0.22078

Number of Clusters	12	13	14	15	16	17	18	19	20
APN	0.18557	0.17232	0.32563	0.32442	0.33371	0.28787	0.22069	0.26755	0.29091
AD	4.00781	3.9394	4.07211	3.97155	3.89618	3.72047	3.57204	3.59344	3.55854
ADM	0.94373	0.80578	1.52803	1.52023	1.53396	1.25453	0.98501	1.21633	1.21699
FOM	0.42805	0.42497	0.417	0.40958	0.39658	0.3843	0.37736	0.37108	0.36569
Connectivity	440.54444	444.14127	448.28373	463.83968	503.24484	476.22222	482.36111	511.56389	515.82698
Dunn	0.02523	0.02523	0.02747	0.02523	0.04669	0.05029	0.04558	0.04695	0.04695
Silhouette	0.20815	0.20608	0.21124	0.21423	0.20601	0.21532	0.21801	0.21499	0.21499

Table 24: Evaluation of the number of clusters, according to APN, AD), ADM, FOM, Connectivity, Dunn and Silhouette validation measures - case study 2.

It can be seen from Fig. 13 that the group with the largest Gap Statistic is the group of only one cluster since with one cluster there is only one way to organize the data. Then it can be seen that the measure decreases to the number of clusters 3, increasing thereafter, with a peak in the number of clusters 5, only surpassed from 14 clusters, increasing considerably thereafter. Although Gap Statistic is larger for larger numbers, since clusters are more likely to end up with less dispersion among themselves (this is the relevant factor for this measure), as mentioned in the previous paragraph, a number of clusters between 4 and 10 is acceptable, the number of clusters 5 was chosen, as it corresponds to the largest Gap Statistic among the 4-10 range. It can also be seen from the Table 24 that the number of clusters 5 is one of the optimal numbers for various measures, between the number of clusters 4 to 10. In this sense, the average value of financial reports in each cluster if uniformly distributed is 183 (917 financial reports/5 clusters). However there will be clusters with different sizes.

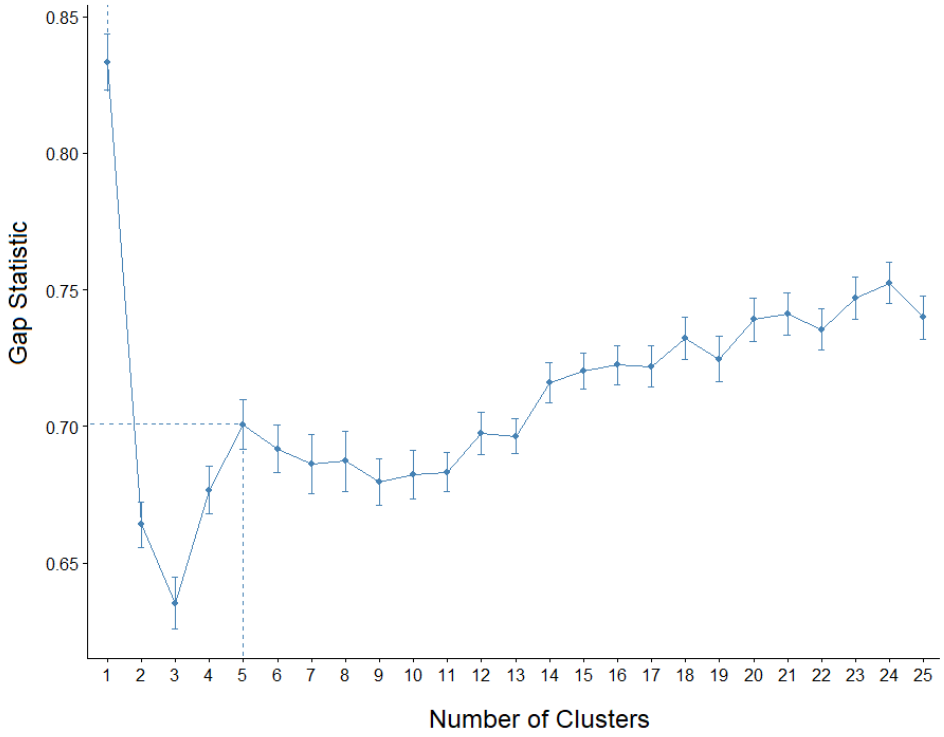


Figure 13: Selection of the number of clusters using the Gap Statistic - case study 2.

4.2.2 *Clustering results - case study 2*

This section presents the results for the 5 readability clusters. For better visualization and discussion, two approaches were made, one focusing on the 43 readability measures, using radar charts (Fig. 15) and another focused on decreasing the complexity of the 43 readability measures, reducing them to only 2 measures, using PCA method (Fig. 14). While radar charts show the results of the 43 readability measures, allowing to discuss and draw accurate conclusions, PCA allows to have a more brief and simpler visualization of the readability clusters.

Analyzing Fig. 14, it can be seen the result of the PCA method, visualizing the result of the clustering method for readability measures in a simpler way. Each of the two variables (Dim1 and Dim2) represents a certain amount of information contained in the 43 readability measures. The proportion of which is represented by the percentage in the axis labels. Dim1 has 68.6% of the information contained in the initial 48 readability measures and Dim2 contains only 22.5%, noting that Dim1 is responsible for more information. It is also noted that 68.6% + 22.5% = 91.1%, meaning that 8.9% (100% - 91.1%) of the information in this graph has been lost, it is natural since the reduction in the size of the data was large, from 48 features to just 2. However, visually it is very worthwhile, since it is clear that at least for the 91.1% of information, the clusters are distinctly different.

Fig. 15 shows the 5 clusters on 4 radar charts (11 readability measures in 3 charts and one chart with 10 measures, 43 in total), allowing to analyze the 4 graphs at the same time. The closer to the outer edge of the chart a measure is, the more readable the cluster is. Ideally, different clusters should be found, each one corresponding to a particular set of readability measures that, as will be assessed in the next section (where external analysis results are analyzed), would match to a specific future crash risk level (e.g., it would be interesting to find a readability cluster that is often associated with a high degree of future crash risk). Thus, in this section, it is first shown and analyzed if the clustering method was effective. Taking this into consideration, it appears that the 5 clusters have obvious differences and there is no cluster that has a similar pattern to others for the 48 readability measures analysed in Fig. 15.

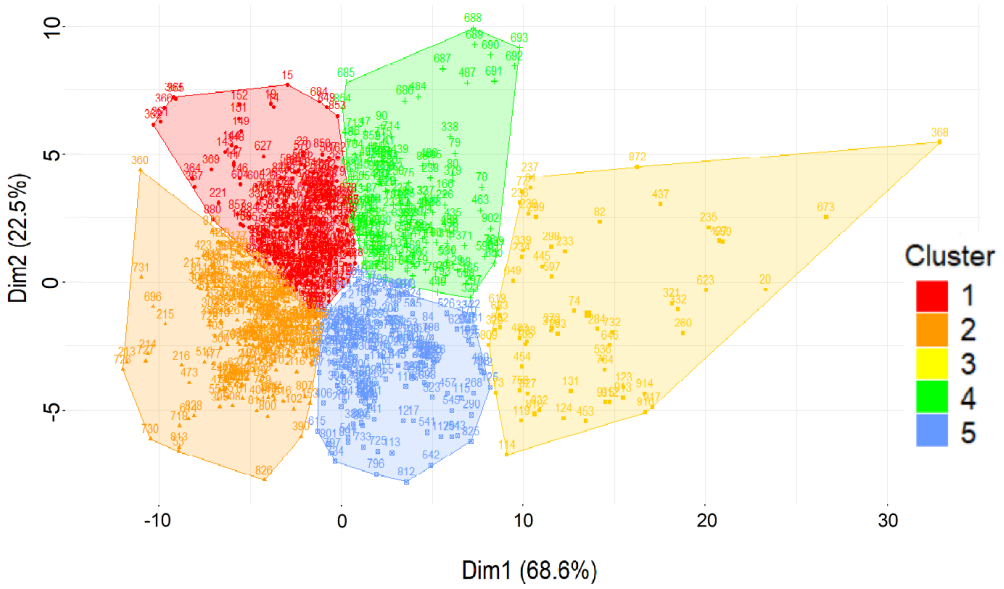


Figure 14: PCA results. The points inside the clusters represent the financial reports - case study 2 (with 5 clusters).

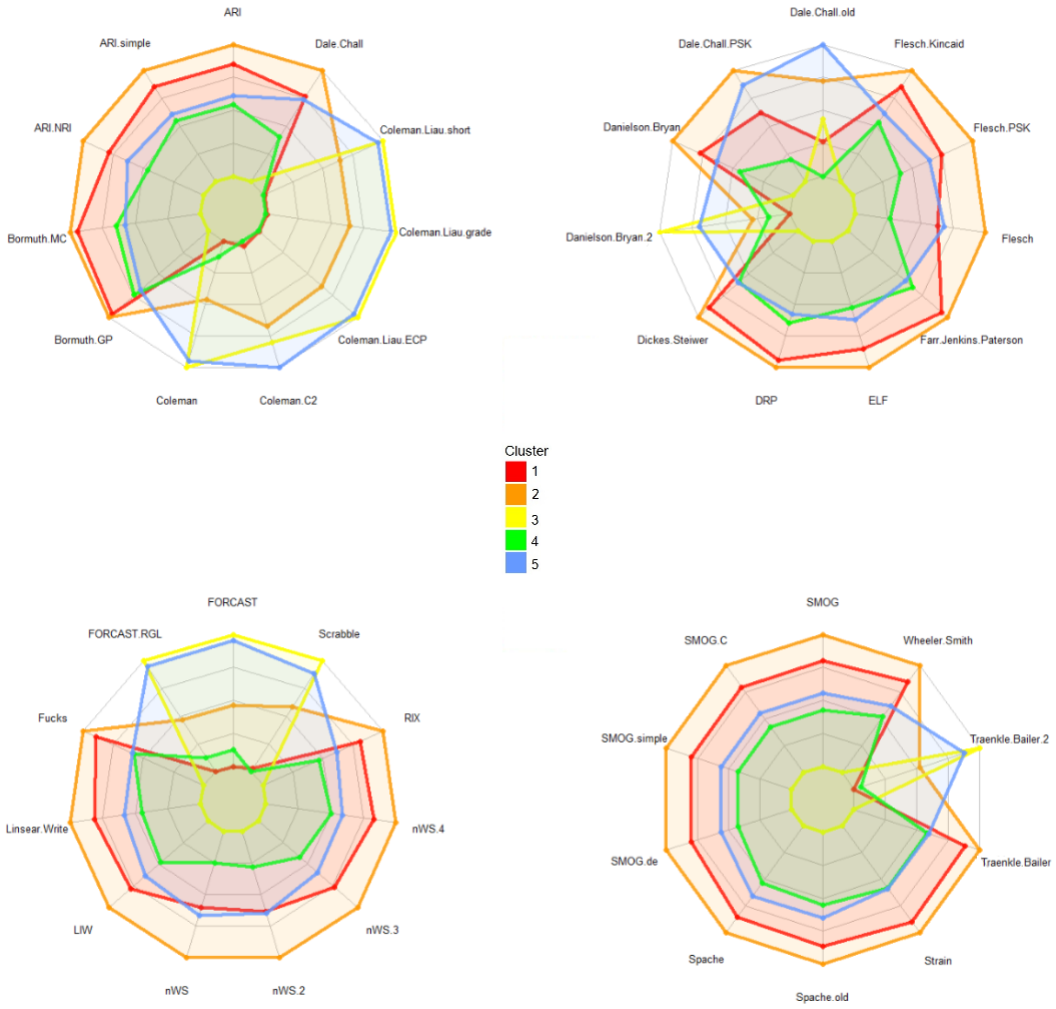


Figure 15: Radar charts with 43 readability measures - case study 2 (with 5 clusters). No FOG, sentence length, or word syllables measures. The closer to the outer edge of the chart a measure is, the more readable the cluster is.

The clusters analyzed in Fig. 15 show a differentiation of them, considering the 43 readability measures visualized in the 4 radar charts of this figure. However, due to the number of features, the view in Fig. 14 becomes more favorable, clearly seeing totally distinct clusters. In this sense, the PCA was of great importance, reducing the initial spectrum from 48 characteristics to only 2, Dim1, containing 68.6% of the initial information of the 48 readability measures and Dim2, containing 22.5% of this information, meaning that Dim1 represents most of the information obtained in this graph. This view proves the success of the k-means clustering method by splitting the clusters properly.

4.2.3 *External analysis results - case study 2*

After observing and analyzing the clusters for the 43 readability measures in Section 4.2.2, it was concluded that it created suitable groups, it is intended to compare these clusters also in terms of future crash risk, externally. Taking this into consideration, it is now possible to analyze the clusters in terms of future crash risk, evaluating future crash risk measures in each cluster (external analysis). Readability metrics measure how readable the financial reports are and the future crash risk allows to evaluate the future company's performance. In this project, future crash risk is obtained from the day following the publication of the financial report until the publication of the next report (next year). Thus, this phase goes through characterizing the readability clusters with future crash risk and realizing what distinguishes them.
As explained, the financial characteristic analyzed in this project is the future crash risk. The financial graphs in Fig. 16, Fig. 17 and Fig. 18 show the future crash risk measures of the 5 clusters analyzing 2 future crash risk measures per graph. The size of the circles is proportional to the size of the clusters. Therefore, it is possible to have the notion of the size of each cluster from these views, also using this feature for their comparison. It is also possible to visualize the correlation between the different future crash risk measures in Table 25 and by cluster in Table 20.
The following points list commonalities between readability and future crash risk measures, taking into account the PCA method.

1. Focusing on the readability clusters depicted in Fig. 14, representing the clusters obtained from readability measures, it can be seen that Cluster 3 is isolated from others, in Dim1. As already explained, (Section 4.2.2 and previous experiment, Section 4.1), the method PCA, represented in this figure (Fig. 14), contains two

	Pearson Correlation
NCSKEW and DUVOL	0.90233
NCSKEW and CRASH_COUNT	0.55599
DUVOL and CRASH_COUNT	0.68632

Table 25: Correlation between future crash risk measures.

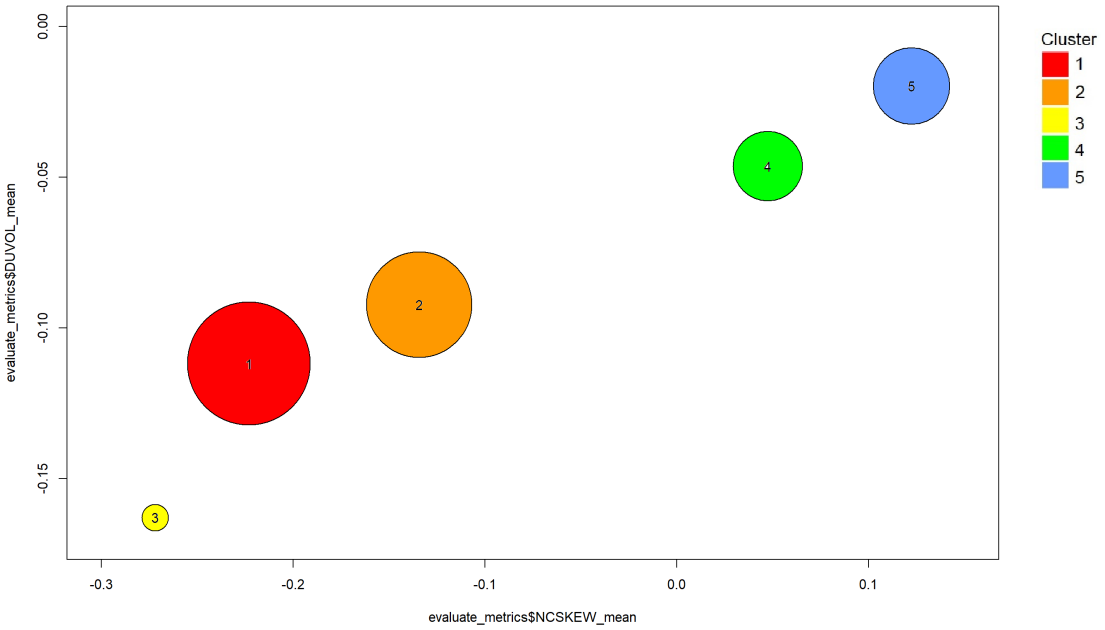


Figure 16: NCSKEW and DUVOL per cluster - case study 2 (with 5 clusters). Circles are proportional to the size of the cluster.

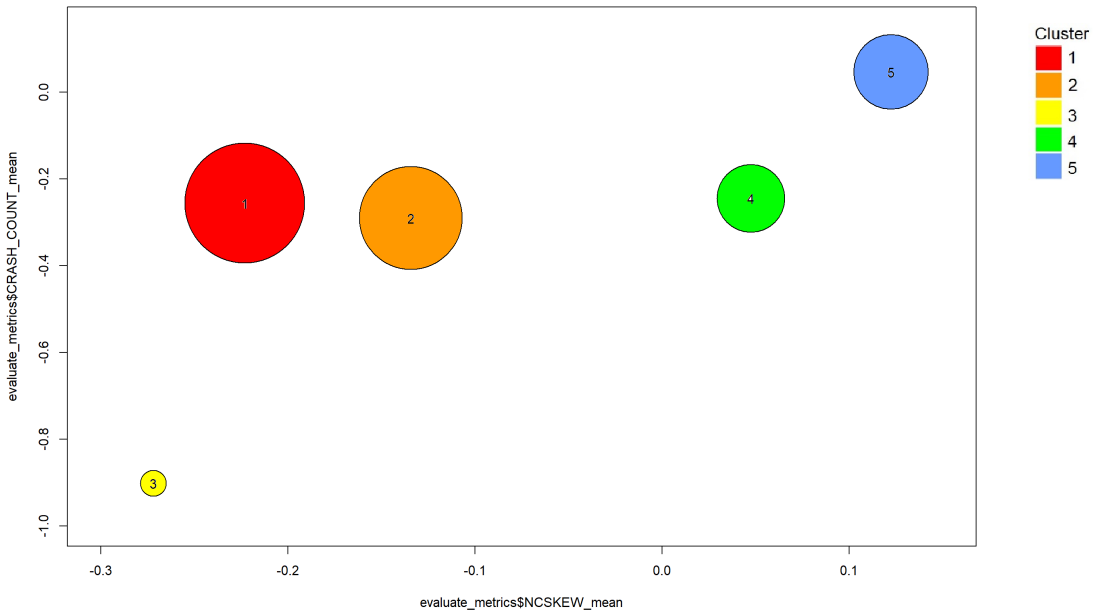


Figure 17: NCSKEW and Crash Count per cluster - case study 2 (with 5 clusters). Circles are proportional to the size of the cluster.

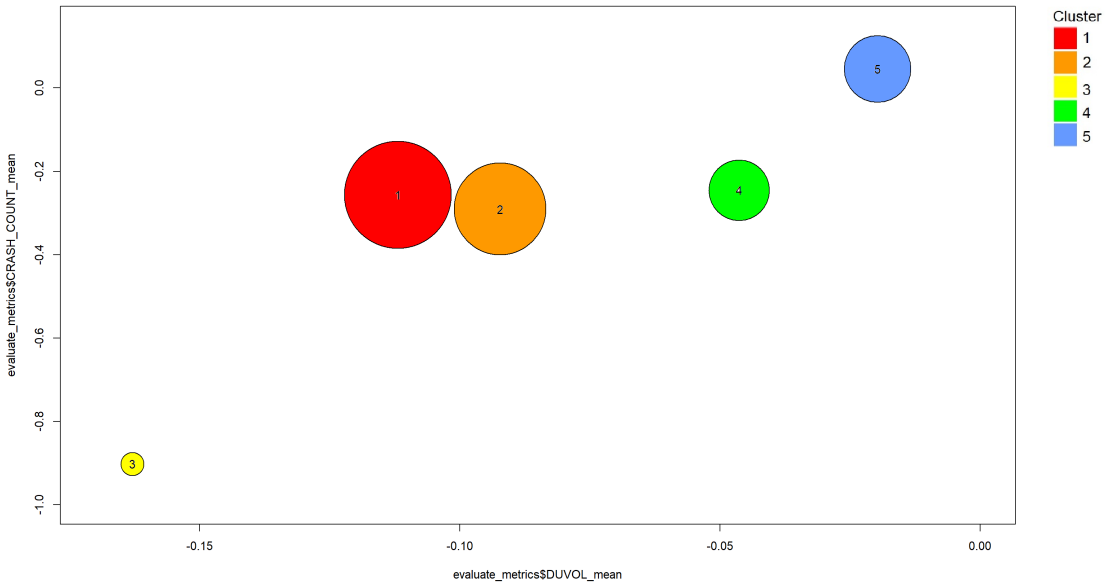


Figure 18: DUVOL and Crash Count per cluster - case study 2 (with 5 clusters). Circles are proportional to the size of the cluster.

features (resulting from the initial 43 readability measures analysed in this experiment), Dim1 and Dim2, as depicted in the image. Dim1 contains 68.6% of the initial information of the 43 readability measures and Dim2 contains 22.5% of the information, meaning that Dim1 represents most of the information obtained in this graph. Therefore, this cluster is isolated in the most of the information, Dim1. Analysing the 3 future crash risk metrics (NCSKEW, DUVOL and Crash Count), in Fig. 16, Fig. 17 and Fig. 18, Cluster 2 always appears further away from all other clusters. Cluster 2 has the lowest future crash risk mean (-0.45, as seen in Table 26) and the lowest NCSKEW, DUVOL and Crash Count. As mentioned, this cluster also has a differentiation in most readability information (68.6 %), as seen in Fig.14, therefore this may indicate a relationship between some readability measures (68.6 % of the information) and future crash risk measures, but it does not show what type of relationship is (how do the measures affect each other).

2. Accordingly, and for the same reasons as the previous paragraph, cluster 4 and cluster 5 are also about the same value as Dim1 in Fig. 14, and these clusters are therefore quite similar (similar in 68.6% of readability information, as explained in the previous point). In Fig. 16, Fig. 17 and Fig. 18, it can be noticed that these clusters are also very close in future crash risk measures, so these clusters are very similar in readability and future crash risk.

Other relationships can be seen by analyzing the cluster rankings for readability measures (these can be grouped into similar groups) and then analyzing the cluster rankings for future crash risk measures. Comparing both rankings is important, since a similar clustering order between several readability measures and some future crash risk measures can be a great indicator of the relationship between these measures. Thus, since 43 readability measures are analysed, it is intended to group the measures in similar groups and considering their cluster order. After this, this groups will be compared with the results of future crash risk rankings by measure. These groups can be compared in terms of readability and future crash risk at the same time for some clusters, as it will be seen. Then, it is presented the readability groups and analyzed the points that these groups have in common considering their calculation formula, using the information from the Table 2 to the Table 9:

1. Group 1 (measures sorted in descending order in Table 28): 37 measures including ARI family (3 metrics), Bormuth family (2 metrics),

	size_clusters	NCSKEW_mean	DUVOL_mean	CRASH_COUNT_mean	mean_all_CR_metrics
Cluster *1*	281	-0.2232	-0.11189	-0.25623	-0.19710
Cluster *2*	241	-0.13429	-0.09226	-0.29046	-0.17233
Cluster *3*	61	-0.27209	-0.16294	-0.90164	-0.44555
Cluster *4*	159	0.04741	-0.04631	-0.24528	-0.08139
Cluster *5*	175	0.12253	-0.0197	0.04571	0.04951

Table 26: Cluster size and external features mean - case study 2 (with 5 clusters)

	NCSKEW and DUVOL	NCSKEW and CRASH_COUNT	DUVOL.and.CRASH_COUNT
Cluster 1	0.90694	0.56385	0.6895
Cluster 2	0.88851	0.50686	0.66838
Cluster 3	0.86028	0.43262	0.59395
Cluster 4	0.92026	0.644	0.75073
Cluster 5	0.90263	0.5385	0.65342

Table 27: Correlation of future crash risk measures per cluster - case study 2 (with 5 clusters)

Coleman.C2, Coleman.Liau.ECP, Coleman.Liau.grade, Coleman.Liau.short, Dale.Chall, Dale.Chall.old, Dale.Chall.PSK, Danielson.Bryan, Dickes.Steiwer, DRP, ELF, Farr.Jenkins.Paterson, Flesch, Flesch.PSK, Flesch.Kincaid, Fucks, Linsear.Write, LIW, nWS, nWS.2, nWS.3, nWS.4, Scrabble, SMOG family (4 metrics), Spache family (2 metrics), Strain, Traenkle.Bailer and Wheeler.Smith.

This group contains major emphasis on Average Sentence Length (ASL) and Average Word Length (AWL), also considering the size of syllables of a sentence, n_{wsy}, (e.g., ELF, Farr.Jenkins.Paterson, Flesch, Flesch.PSK), with different variations (e.g., word count with one, two or three syllables per word, depending on the metric). Less frequently, another components are also used as: number of "difficult" words not matching the Dale-Chall list of "familiar" words, n_{wd} (e.g., Dale.Chall.PSK), number of words, n_w (e.g., Dale.Chall. PSK, Flesch, Flesch.PSK), number of characters, n_c (e.g., Danielson.Bryan), the number of blanks, n_{blank} (e.g., Danielson.Bryan) and number of sentences, n_{st} (e.g., Danielson.Bryan, ELF). See notation in Section 2.2.4.

2. Group 2 (measures sorted in descending order in Table 29): 10 measures including Coleman family (5 measures), Danielson.Bryan.2, FORCAST family (2 metrics), Scrabble and Traenkle.Bailer.2.

This group does not use Average Sentence Length (ASL), unlike previous groups. This group is based on n_{wsy} (e.g., Coleman, Coleman.C2, FORCAST), n_w (e.g., Traenkle.Bailer2, Coleman.Liau.ECP, FORCAST), n_{st} (e.g., Danielson.Bryan2). Other less commonly used components are the number of prepositions, n_{conj} 3(e.g., Traenkle.Bailer2) and the number of conjunctions, n_{nconj} (e.g., Traenkle.Bailer2). See notation in Section 2.2.4.

3. Group 3 (measures sorted in descending order in Table 30): 5 measures including Bormuth family (2 measures), DRP, Dickes.Steiwer and Farr.Jenkins.Paterson.

   These measures have the same main characteristics as Group 1 measures. However, since the components are used differently, these metrics have different results in this case study, in one of the cases that will be analyzed.

Looking at Table 26, it is now possible to ranking the clusters regarding the future crash risk metrics: The three future crash risk metrics have some consistency in the order they define clusters, since they are partially correlated, as shown in Fig. 27. However, Crash Count has a slightly different order (in cluster 1 and 2). The cluster rankings of the future crash risk measures and their mean are shown:

1. Sorting clusters by NCSKEW mean, ascending form: [3,1,2,4,5]. Color order: [yellow, red, orange, green, blue].
2. Sorting clusters by DUVOL mean, ascending form: [3,1,2,4,5]. Color order: [yellow, red, orange, green, blue].
3. Sorting clusters by Crash Count mean, ascending form: [3,2,1,4,5]. Color order: [yellow, orange, red, green, blue].

To show the evidence of the relationship of these readability measures groups with future crash risk measures, it will be analyzed by the order of clusters for the future crash risk measures mentioned above and their relationship with the proposed groups. As already listed, the ascending order of clusters with respect to future crash risk measures is again mentioned: (NCSKEW: [3,1,2,4,5]), (DUVOL: [3,1,2,4,5]) and (Crash Count: [3,2,1,4,5]). These sortings will be used as follows, analysing the relationship of readability measures groups and future crash risk measures:

1. Part of Group 2 (Coleman.C2, Coleman.Liau.ECP, Coleman.Liau.grade, Coleman.Liau.short and Scrabble) with Crash Count: Considering the clusters order

	mean_all_scale	ARI	ARI.simple	ARI.NRI	Bormuth.MC	Bormuth.GP	Coleman.C2
Cluster 2	0.903818981212074	-16.7512242902	-74.1407056435331	-14.9519561823674	-4.90791079514032	-317833396.599133	29.7727108403473
Cluster 1	0.396956189885797	-17.6268723502758	-75.8441373557156	-15.9068397699208	-5.04793540703877	-345859839.311022	27.659193533484
Cluster 5	-0.424284332365531	-19.0456343554397	-78.768574466159	-16.5799683277227	-5.95769808030393	-566879127.938654	30.8677624725863
Cluster 4	-0.647859029205778	-19.4318784296159	-79.4627517348865	-17.3051299759275	-5.78145744186998	-518070882.83064	27.5212798296548
Cluster 3	-2.49353639380871	-22.6204411560521	-85.9337184601294	-19.3572808239076	-7.39568754954055	-1092843021.83268	30.2093970628109

	Coleman.Liau.ECP	Coleman.Liau.grade	Coleman.Liau.short	Dale.Chall	Dale.Chall.old	Dale.Chall.PSK	Danielson.Bryan
Cluster 2	32.4116732469947	-14.1830218836305	-14.1841290246686	5.05500988805453	-11.7069402865609	-9.78922327318302	-7.59805327345514
Cluster 1	29.8655002329838	-14.8806834741615	-14.8818193553696	3.72173741054028	-11.8844316395526	-9.93485856233233	-7.84538150253667
Cluster 5	33.7056934753611	-13.8284551649772	-13.8296311131297	3.55350995354828	-11.6008387263788	-9.83904115284935	-8.00484462251674
Cluster 4	29.8122741249179	-14.895267640676	-14.896458527665	1.62744359826627	-11.9850094661213	-10.0971090020247	-8.21092430171681
Cluster 3	33.8682680277124	-13.7839090873347	-13.7851538735821	-0.687491806311862	-11.8177325850341	-10.1725367118464	-8.69251744162087

	Dickes.Steiwer	DRP	ELF	Farr.Jenkins.Paterson	Flesch	Flesch.PSK	Flesch.Kincaid
Cluster 2	-487.268501906187	-590.791079514032	-12.314018118133	-57.5905966680324	25.7300595054872	-8.15320366268472	-16.2617711293443
Cluster 1	-504.083486520699	-604.793540703878	-13.0885839470271	-58.3057612512232	21.9556696861542	-8.37193126302863	-16.9565986028307
Cluster 5	-547.906651247474	-695.769808030393	-14.2848373075211	-63.1088520869358	22.4481296194493	-8.45655219150523	-18.077100066071
Cluster 4	-549.595161374439	-678.145744186998	-14.7807862932817	-62.1583448425356	18.1535045745845	-8.66469052309355	-18.4316825225759
Cluster 3	-638.858028579711	-839.568754954056	-17.5609714918325	-70.7164144982077	15.4275210065973	-9.00817093178139	-20.9190308839037

	Fucks	Linsear.Write	LIW	nWS	nWS.2	nWS.3	nWS.4
Cluster 2	-140.28077175184	-25.727621272035	-60.7601169179304	-11.9092778231108	-12.274341357734	-11.1307374585646	-12.0132015453439
Cluster 1	-146.94127349088	-27.8249109637884	-63.1951456006943	-12.5757936557005	-12.9094584503125	-11.6563381753554	-12.5603975464401
Cluster 5	-166.218293164138	-30.4299736146278	-64.8203570116465	-12.4698993062539	-12.8830263292781	-12.0332759402996	-13.3363682260649
Cluster 4	-166.879409784485	-31.9316956137474	-66.6136595593328	-13.1697830991271	-13.5201454250369	-12.4302218875276	-13.616229757816
Cluster 3	-203.641651771504	-36.9783519160731	-71.8940950073065	-13.598135782398	-14.018862732702	-13.3690902987795	-15.2421177258849

	Scrabble	SMOG	SMOG.C	SMOG.simple	SMOG.de	Spache	Spache.old
Cluster 2	1.61695883825924	-17.6113502101116	-16.9240382424727	-16.8851871621396	-11.8851871621396	-8.28464549357756	-9.21066367935265
Cluster 1	1.60708097454722	-18.191143168049	-17.4723267815942	-17.4410768629425	-12.4410768629425	-8.46144718336637	-9.40564468688926
Cluster 5	1.62229641297061	-18.8812357438216	-18.125572872742	-18.1027188339613	-13.1027188339613	-8.70114585074658	-9.72451903256985
Cluster 4	1.60657344204716	-19.2650602364916	-18.4891889388807	-18.4707193063199	-13.4707193063199	-8.85309925743331	-9.87001402692001
Cluster 3	1.62439854731758	-20.4874966153601	-19.6484724420417	-19.6427580204794	-14.6427580204794	-9.50407826326308	-10.672321519741

	Strain	Traenkle.Bailer	Wheeler.Smith
Cluster 2	-14.4994649230366	-545.210476167318	-123.14018118133
Cluster 1	-15.1651917028954	-562.111170685785	-130.885839470271
Cluster 5	-17.2294588947057	-605.710411527726	-142.848373075211
Cluster 4	-17.2803638238101	-607.457159684075	-147.807862932817
Cluster 3	-21.1735715276141	-695.147126803352	-175.609714918325

Table 28: Readability measures Group 1 for case study 2 (with 5 clusters)

	mean_all_scale	Coleman	Coleman.C2	Coleman.Liau.ECP	Coleman.Liau.grade	Coleman.Liau.short
Cluster 3	1.19030347447876	33.149111088197	30.2093970628109	33.8682680277124	-13.7839090873347	-13.7851538735821
Cluster 5	1.09487710316762	32.9196068716177	30.8677624725863	33.7056934753611	-13.8284551649772	-13.8296311131297
Cluster 2	0.144214640771086	30.6332258569585	29.7727108403473	32.4116732469947	-14.1830218836305	-14.1841290246686
Cluster 4	-0.520766925434813	29.0339187545841	27.5212798296548	29.8122741249179	-14.895267640676	-14.896458527665
Cluster 1	-0.769273282132506	28.4362510876149	27.659193533484	29.8655002329838	-14.8806834741615	-14.8818193553696

	Danielson.Bryan.2	FORCAST	FORCAST.RGL	Scrabble	Traenkle.Bailer.2
Cluster 3	81.5319540718303	-11.6745219664887	-11.2719741631376	1.62439854731758	-307.223023133906
Cluster 5	80.4225375360597	-11.7012085033003	-11.3013293536303	1.62229641297061	-309.535666241896
Cluster 2	78.9553839572913	-11.9670667608188	-11.5937734369007	1.61695883825924	-316.538257838831
Cluster 4	78.502823030841	-12.1530327029553	-11.7983359732509	1.60657344204716	-325.784244485177
Cluster 1	77.9034527375872	-12.2225289433006	-11.8747818376306	1.60708097454722	-326.896623996348

Table 29: Readability measures Group 2 for case study 2 (with 5 clusters)

	mean_all_scale	Bormuth.MC	Bormuth.GP	DRP	Dickes.Steiwer	Farr.Jenkins.Paterson
Cluster 2	0.73456479879323	-4.90791079514032	-317833396.599133	-590.791079514032	-487.268501906187	-57.5905966680324
Cluster 1	0.550958529793262	-5.04793540703877	-345859839.311022	-604.793540703878	-504.083486520699	-58.3057612512232
Cluster 4	-0.410866982298068	-5.78145744186998	-518070882.83064	-678.145744186998	-549.595161374439	-62.1583448425356
Cluster 5	-0.641961256585344	-5.95769808030393	-566879127.938654	-695.769808030393	-547.906651247474	-63.1088520869358
Cluster 3	-2.52751464415159	-7.39568754954055	-1092843021.83268	-839.568754954056	-638.858028579711	-70.7164144982077

Table 30: Readability measures Group 3 for case study 2 (with 5 clusters)

[3,2,1,4] (without cluster 5), painted [yellow, orange, red and green], these clusters follow this order for readability metrics Group 3 (in descendent order) and the future crash risk metric Crash Count (in ascending order).

2. Group 3 with Crash Count: Considering the clusters order [2,1,4,5] (without cluster 3), painted [orange, red, green and blue], these clusters follow this order for readability metrics Group 3 (in descendent order) and the future crash risk metric Crash Count (in ascending order).

3. Group 1 with Crash Count: Considering the clusters order [2,1,4] (without cluster 3 and 5), painted [orange, red and green], these clusters follow this order for readability metrics Group 1 (in descendent order) and the future crash risk metric Crash Count (in ascending order).

4. Group 2 with NCSKEW: Considering the clusters order [3,2,4] (without clusters 1 and 5), painted [yellow, orange and green], these clusters follow this order for readability metrics Group 2 (in descendent order) and the future crash risk metric NCSKEW (in ascending order).

5. Group 2 with DUVOL: Considering the clusters order [3,2,4] (without cluster 1 and 5), painted [yellow, orange and green], these clusters follow this order for readability metrics Group 2 (in descendent order) and the future crash risk metric DUVOL (in ascending order).

6. Group 2 with Crash Count: Considering the clusters order [3,2,4] (without cluster 1 and 5), painted [yellow, orange and green], these clusters follow this order for readability metrics Group 2 (in descendent order) and the future crash risk metric Crash Count (in ascending order).

7. Group 3 with NCSKEW: Considering the clusters order [2,4,5] (without cluster 1 and 3), painted [orange, green and blue], these clusters follow this order for readability metrics Group 3 (in descendent order) and the future crash risk metric NCSKEW (in ascending order).

8. Group 3 with DUVOL : Considering the clusters order [2,4,5] (without cluster 1 and 3), painted [orange, green and blue], these clusters follow this order for readability metrics Group 3 (in descendent order) and the future crash risk metric DUVOL (in ascending order).

9. Group 3 with Crash Count: Considering the clusters order [2,4,5] (without cluster 1 and 3), painted [orange, green and blue], these clusters follow this order for readability metrics Group 3 (in descendent order) and the future crash risk metric Crash Count (in ascending order).

10. Group 3 with Crash Count: Considering the clusters order [1,4,5] (without cluster 2 and 3), painted [red, green and blue], these clusters follow this order for readability metrics Group 3 (in descendent order) and the future crash risk metric Crash Count (in ascending order).

4.3 DISCUSSION

The results achieved were analyzed in Section 4.1 and Section 4.2, where two different case studies are evaluated. As explained, while the first case study analyzes all the different readability measures exposed in Section 2.2.4, the second case study analyzes only 43 readability measures, excluding the FOG family, sentence length and word syllables from all 48 measures of case study 1. In case study 1, the optimal number of clusters was 6, and in the case study 2, the optimal number of clusters was 5, allowing the financial reports to be divided into two different ways, one with 6 clusters and another with 5 clusters. There were no significant differences between the two main clustering approaches. Therefore, there seems to be some redundancy in the readability data features, which translates into consistent clustering results when using different sets of features. This suggests that the FOG family, sentence length and word syllables metrics can be replaced by the other readability measures. Summarizing and discussing the similarities between the readability metrics and future crash risk metrics in the two case studies, it can be highlighted:

1. In both case studies, it is possible to find a readability cluster farther from the others, by analyzing the results of the PCA method, in Fig. 8 (case study 1) and Fig. 14 (case study 2). When analyzed externally, this cluster (cluster 2 in case study 1 and cluster 3 in case study 2) has the smallest NCSKEW, DUVOL and Crash Count (as it can be seen in Table 19 and Table 26). Thus, there appears to be a relationship between a set of specific readability values with lower future crash risk metrics.

2. Analyzing the results of the PCA method, it was also evidenced that there are some similarities in the feature Dim1 (which represents 68.6% of readability in-

formation, as shown in fig 8 and 14) for some clusters. This similarity also corresponds in future crash risk metrics, in the same clusters. One example was given for case study 1 (clusters 1 and 3) and two examples for case study 2 (clusters 1 and 2 and clusters 4 and 5) with emphasis on the Crash Count metric in the first example (this metric is almost the same in cluster 1 and 3). Taking this into consideration, it can be said that there may be a possible relationship in 68.6% of readability information, which may mean that a specific set of readability metrics and future crash risk measures can have a relationship.

Analyzing the readability measures in both case studies and their groups, considering the rankings (Fig. 21, Fig. 22 and Fig. 23 for case study 1 and Fig. 28, Fig. 29 and Fig. 30 for case study 2), as well as the mentioned future crash risk rankings (case study 1 - NCSKEW: [2,1,3,6,4,5], DUVOL: [2,1,6,5,4,3] and Crash Count: [2,5,6,3,1,4]; case study 2 - NCSKEW: [3,1,2,4,5], DUVOL: [3,1,2,4,5]) and Crash Count: [3,2,1,4,5]), some relationships were documented in Section 4.1.3 and Section 4.2.3. The analysis consisted in creating similar groups of readability measures, in each case study (each group has a set of readability measures with the same ranking). After that, the clusters were also sorted by future crash risk metrics (NCSKEW, DUVOL and Crash Count, having 3 rankings). In several cases, the readability measures rankings match the future crash risk measures rankings, in symmetrical way. In symmetrical way means, for example, that the most readable cluster of the ranked clusters is the least crash-prone cluster and the less readable cluster is the one with higher future crash risk. In this sense, the documented evidence were, briefly:

1. Case study 1 - Group 3 with NCSKEW: Considering the clusters order [2,6,4,5].
2. Case study 1 - Group 3 with DUVOL: Considering the clusters order [2,6,4,3].
3. Case study 1 - Group 3 with Crash Count: Considering the clusters order [2,6,3,1].
4. Case study 1 - Group 2 with NCSKEW: Considering the clusters order [1,6,5].
5. Case study 1 - Group 2 with DUVOL: Considering the clusters order [1,6,5].
6. Case study 1 - Group 1 with Crash Count: Considering the clusters order [3,1,4].
7. Case study 2 - Part of Group 2 (Coleman.C2, Coleman.Liau.ECP, Coleman.Liau.grade, Coleman.Liau.short and Scrabble) with Crash Count: Considering the clusters order [3,2,1,4].

8. Case study 2 - Group 3 with Crash Count: Considering the clusters order [2,1,4,5].
9. Case study 2 - Group 1 with Crash Count: Considering the clusters order [2,1,4].
10. Case study 2 - Group 2 with NCSKEW: Considering the clusters order [3,2,4].
11. Case study 2 - Group 2 with DUVOL: Considering the clusters order [3,2,4].
12. Case study 2 - Group 2 with Crash Count: Considering the clusters order [3,2,4].
13. Case study 2 - Group 3 with NCSKEW: Considering the clusters order [2,4,5].
14. Case study 2 - Group 3 with DUVOL: Considering the clusters order [2,4,5].
15. Case study 2 - Group 3 with Crash Count: Considering the clusters order [2,4,5].
16. Case study 2 - Group 3 with Crash Count: Considering the clusters order [1,4,5].

These relationships demonstrate several points of contact between readability and future crash risk measures. We also discuss the possibility of complementary evidences, such as a group of readability measures predicting part of each crash risk metric. For example, Group 3 metrics of study case 1 have the ability to be related to all clusters (4 different clusters in each relation with a future crash risk measure). This group has 11 metrics including Coleman family (5 metrics), Danielson.Bryan.2, FORCAST family (2 metrics), Scrabble, Traenkle.Bailer.2 and meanWordSyllables. These metrics are correlated to the 3 future crash risk measures and include all 6 clusters (4 at a time):

1. NCSKEW: follows the same ranking (symmetrically) as the referred readability metrics for clusters [2,6,4,5].
2. DUVOL: follows the same ranking (symmetrically) as the referred readability metrics for clusters [2,6,4,3].
3. Crash Count: follows the same ranking (symmetrically) as the referred readability metrics for clusters [2,6,3,1].

Thus, in Group 3 of case study 1, all clusters are involved in at least one relationship with the three future crash risk metrics. In this case, all future crash risk metrics established different rankings so it would be impossible for a readability group to match the three future crash risk metrics. Even so, the readability metrics of this

group are related to all future crash risk metrics, showing the possibility of a strong relationship between these readability metrics and the future crash risk metrics.
It is also noted that the mentioned readability metrics of Group 3 for case study 1, which have points of contact with all future crash risk metrics, have in common not using Average Sentence Length (ASL). The metrics of this group are based on n_{wsy} (e.g., Coleman, Coleman.C2, FORCAST), n_w (e.g., Traenkle.Bailer2, Coleman.Liau.ECP, FORCAST), n_{st} (e.g., Danielson.Bryan2). Other less commonly used components are the number of prepositions, n_{conj} (e.g., Traenkle.Bailer2) and the number of conjunctions, n_{nconj} (e.g., Traenkle.Bailer2). See notation in Section 2.2.4. These components can be extremely important as they pertain to the readability measures (Group 3) in which we find the most correlation with future crash risk metrics. Thus, these components may be the most important components in a financial report to predict the financial results of the company.

5

CONCLUSION

5.1 CONCLUSIONS

Referring to the objectives initially proposed in the Section 1.2, it is possible to say that the objectives were achieved with satisfactory results. The following are the main conclusions of this study:

1. The first case study analyzes all the different readability measures exposed in Section 2.2.4 and the second case study analyzes only 43 readability measures, excluding the FOG family, sentence length and word syllables from all 48 measures of case study 1. There were no significant differences between the two main clustering approaches. Therefore, there seems to be some redundancy in the readability data features, which translates into consistent clustering results when using different sets of features. In this sense, with the data analyzed in this study, it can be concluded that the FOG family, sentence length and word syllables metrics can be replaced by the other readability measures.

2. The readability results of the PCA method (Fig. 8 and Fig. 14) allow us to conclude that in 68.6% of the readability information (corresponding to one dimension), there are some relationships between this readability information and future crash risk (exposed in Table 19 and Table 26), such as the same cluster isolated from the others in readability and crash risk results, as well as some correspondent similarities in both results. In this sense, for the analyzed data, it is possible to conclude that there are evidences of relationships between 68.6% of the readability information and the crash risk metrics.

3. Analysing all readability metrics, in Fig. 9 (case study 1) and Fig. 15 (case study 2), there are several relationships between some readability measures and future crash risk metrics (exposed in Table 19 and Table 26). It is possible to

highlight Coleman family (5 metrics), Danielson.Bryan.2, FORCAST family (2 metrics), Scrabble and Traenkle.Bailer.2 metrics (Group 3 of case study 1 and Group 2 of case study 2), which in addition to having several relationships with future crash risk metrics, they include all the future crash risk metrics (NCSKEW, DUVOL and Crash Count). These metrics point to a more readable financial reporting ratio being related to less crash-prone companies.

4. The previous point suggests that Coleman family (5 metrics), Danielson.Bryan.2, FORCAST family (2 metrics), Scrabble and Traenkle.Bailer.2 metrics are related to future crash risk metrics. In this sense, and since these metrics share similar components, these components may be related to future crash risk. These metrics have in common not using Average Sentence Length (ASL) and they are based on n_{wsy} (e.g., Coleman, Coleman.C2, FORCAST), n_w (e.g., Traenkle.Bailer2, Coleman.Liau.ECP, FORCAST) and n_{st} (e.g., Danielson.Bryan2). Other less commonly used components are the number of prepositions, n_{conj} (e.g., Traenkle.Bailer2) and the number of conjunctions, n_{nconj} (e.g., Traenkle.Bailer2). See notation in Section 2.2.4. This study, using S&P 100 from 2008 to 2017, suggests that these components may be more related to crash risk metrics NCSKEW, DUVOL and Crash Count.

5.2 PROSPECT FOR FUTURE WORK

This work allowed us to understand some relationships between readability and crash risk metrics, highlighting some metrics and components that may be more related to crash risk. With this in mind, the future work has several directions that can be followed:

1. Retry clustering with the most related metrics: Coleman family (5 metrics), Danielson.Bryan.2, FORCAST family (2 metrics), Scrabble and Traenkle.Bailer.2 metrics. Understand more details of these metrics and why they might be linked to future crash risk metrics. Test if they are related to other financial indicators.
2. Expand data size. For example, using the S&P 500 index or analysing more than 10 years.
3. Test other financial indicators, in addition to future crash risk.
4. Finding other readability metrics or components that are related to future crash risk.

5. With the knowledge gathered, develop new readability measures, that may be more related to companies' financial performance..

6. Develop prediction algorithms, building various Machine Learning models using different readability measures to predict different indicators of financial performance (one in each model). In this process, several models are created, where each model predicts a different financial indicator. Analyze the results, seeking conclusions on a set of metrics that can predict a financial indicator.

BIBLIOGRAPHY

Mehdi Allahyari, Seyed Amin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. A brief survey of text mining: Classification, clustering and extraction techniques. *CoRR*, abs/1707.02919, 2017. URL http://arxiv.org/abs/1707.02919.

Jonathan Anderson. Lix and rix: Variations on a little-known readability index. *Journal of Reading*, 26(6):490–496, 1983.

Adam Atkins, Mahesan Niranjan, and Enrico Gerding. Financial news predicts stock market volatility better than close price. *The Journal of Finance and Data Science*, 4(2): 120–137, 2018.

Richard Bamberger and Erich Vanecek. *Lesen, verstehen, lernen, schreiben: die Schwierigkeitsstufen von Texten in deutscher Sprache*. Jugend und Volk, 1984.

BarChart. Provider of real-time or delayed intraday stock and commodities charts and quotes, 2019. URL https://www.barchart.com.

David S Bates. Us stock market crash risk, 1926–2010. *Journal of Financial Economics*, 105(2):229–259, 2012.

CH Björnsson. Läsbarhet, liber. *Stockholm, Sweden*, 1968.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

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

John R Bormuth. Cloze test readability: Criterion reference scores. *Journal of educational measurement*, 5(3):189–196, 1968.

John R Bormuth. Development of readability analysis. 1969.

Jeffrey L Callen and Xiaohua Fang. Religion and stock price crash risk. *Journal of Financial and Quantitative Analysis*, 50(1-2):169–195, 2015.

John S Caylor and Thomas G Sticht. Development of a simple readability index for job reading material. 1973.

Soumen Chakrabarti, Byron Dom, Rakesh Agrawal, and Prabhakar Raghavan. Using taxonomy, discriminants, and signatures for navigating in text databases. In *VLDB*, volume 97, pages 446–455, 1997.

Jeanne Sternlicht Chall and Edgar Dale. *Readability revisited: The new Dale-Chall readability formula*. Brookline Books, 1995.

Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, Rudiger Wirth, et al. Crisp-dm 1.0: Step-by-step data mining guide. *SPSS inc*, 16, 2000.

Joseph Chen, Harrison Hong, and Jeremy C Stein. Forecasting crashes: Trading volume, past returns, and conditional skewness in stock prices. *Journal of financial Economics*, 61(3):345–381, 2001.

Edmund B Coleman. Developing a technology of written instruction: Some determiners of the complexity of prose. *Verbal learning research and the technology of written instruction*, pages 155–204, 1971.

Meri Coleman and Ta Lin Liau. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283, 1975.

CRSP. Center for research in security prices, 2019. URL http://www.crsp.com/.

Edgar Dale and Jeanne S Chall. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54, 1948.

Wayne A Danielson and Sam Dunn Bryan. Computer automation of two readability formulas. *Journalism Quarterly*, 40(2):201–206, 1963.

Alice Davison and Robert N Kantor. On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading research quarterly*, pages 187–209, 1982.

Gus De Franco, Ole-Kristian Hope, Dushyantkumar Vyas, and Yibin Zhou. Analyst report readability. *Contemporary Accounting Research*, 32(1):76–104, 2015.

Paul Dickes and Laure Steiwer. Ausarbeitung von lesbarkeitsformeln für die deutsche sprache. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 9(1):20–28, 1977.

William H DuBay. The principles of readability. *Online Submission*, 2004.

Irving E Fang. The "easy listening formula". *Journal of Broadcasting & Electronic Media*, 11(1):63–68, 1966.

James N Farr, James J Jenkins, and Donald G Paterson. Simplification of flesch reading ease formula. *Journal of applied psychology*, 35(5):333, 1951.

Rudolph Flesch. A new readability yardstick. *Journal of applied psychology*, 32(3):221, 1948.

Wilhelm Fucks. *Unterschied des Prosastils von Dichtern und anderen Schriftstellern: ein Beispiel mathematischer Stilanalyse*. Bouvier, 1955.

John Gantz and David Reinsel. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView: IDC Analyze the future*, 2007 (2012):1–16, 2012.

Robert Gunning. The technique of clear writing. 1952. *New York, NYMcGraw-Hill*, 1952.

Vishal Gupta, Gurpreet S Lehal, et al. A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence*, 1(1):60–76, 2009.

Eui-Hong Sam Han, George Karypis, and Vipin Kumar. Text categorization using weight adjusted k-nearest neighbor classification. In *Pacific-asia conference on knowledge discovery and data mining*, pages 53–65. Springer, 2001.

Amy P Hutton, Alan J Marcus, and Hassan Tehranian. Opaque financial reports, r2, and crash risk. *Journal of financial Economics*, 94(1):67–86, 2009.

Investor.gov. Form 10-k, 2019. URL https://www.investor.gov/additional-resources/general-resources/glossary/form-10-k.

Ishares. Sp 100 etf, 2019. URL https://www.ishares.com/us/products/239723/ishares-sp-100-etf.

Anders Johansen, Didier Sornette, and Olivier Ledoit. Predicting financial crashes using discrete scale invariance. *arXiv preprint cond-mat/9903321*, 1999.

Kenneth. Kenneth r. french site, 2019. URL http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

Jeong-Bon Kim and Liandong Zhang. Accounting conservatism and stock price crash risk: Firm-level evidence. *Contemporary Accounting Research*, 33(1):412–441, 2016.

Jeong-Bon Kim, Yinghua Li, and Liandong Zhang. Corporate tax avoidance and stock price crash risk: Firm-level analysis. *Journal of Financial Economics*, 100(3):639–662, 2011.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. 1975.

George R Klare. Assessing readability. *Reading research quarterly*, pages 62–102, 1974.

George Roger Klare et al. Measurement of readability. 1963.

Genell L Knatterud, Frank W Rockhold, Stephen L George, Franca B Barton, CE Davis, William R Fairweather, Tom Honohan, Richard Mowery, and Robert O'Neill. Guidelines for quality assurance in multicenter trials: a position paper. *Controlled clinical trials*, 19(5):477–493, 1998.

Deanne Larson and Victor Chang. A review and future direction of agile, business intelligence, analytics and data science. *International Journal of Information Management*, 36(5):700–710, 2016.

Alastair Lawrence. Individual investors and financial disclosure. *Journal of Accounting and Economics*, 56(1):130–147, 2013.

Reuven Lehavy, Feng Li, and Kenneth Merkley. The effect of annual report readability on analyst following and the properties of their earnings forecasts. *The Accounting Review*, 86(3):1087–1115, 2011.

Feng Li. Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and economics*, 45(2-3):221–247, 2008.

Tim Loughran and Bill McDonald. Measuring readability in financial disclosures. *The Journal of Finance*, 69(4):1643–1671, 2014.

G Harry Mc Laughlin. Smog grading-a new readability formula. *Journal of reading*, 12 (8):639–646, 1969.

Bill McDonald. Bill mcdonald, 2019. URL https://sraf.nd.edu/data/stage-one-10-x-parse-data/.

Marlene M Most, Shirley Craddick, Staci Crawford, Susan Redican, Donna Rhodes, Fran Rukenbrod, Reesa Laws, Dash-Sodium Collaborative Research Group, et al. Dietary quality assurance processes of the dash-sodium controlled diet study. *Journal of the American Dietetic Association*, 103(10):1339–1346, 2003.

Azadeh Nikfarjam, Abeed Sarker, Karen O'connor, Rachel Ginn, and Graciela Gonzalez. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681, 2015.

Cathy O'Neil and Rachel Schutt. *Doing data science: Straight talk from the frontline.* " O'Reilly Media, Inc.", 2013.

Richard D Powers, William A Sumner, and Bryant E Kearl. A recalculation of four adult readability formulas. *Journal of Educational Psychology*, 49(2):99, 1958.

Daniel Ramage, Susan T. Dumais, and Daniel J. Liebling. Characterizing microblogs with topic models. In *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010*, 2010.

SEC. *Plain English Handbook*. Office of Investor Education and Assistance U.S. Securities and Exchange Commission, 1998.

sec. Form 10-q, 2019a. URL https://www.sec.gov/fast-answers/answersform10qhtm.html.

sec. Changeover to the sec's new smaller reporting company system by small business issuers and non-accelerated filer companies, 2019b. URL https://www.sec.gov/info/smallbus/secg/smrepcosysguid.pdf.

SEC.gov. Sec.gov, 2019. URL https://www.sec.gov/edgar/searchedgar/companysearch.html.

John Seely. *Oxford AZ of grammar and punctuation*. Oxford University Press, 2013.

Jyoti Sharma, Mrs Shashi Sharma, and Ruchi Pandey. A complete review of concept of data mining. *International Journal For Technological Research In Engineering*, 5(6): 3143–3146, 2018.

Edgar A Smith and RJ Senter. Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories (US)*, pages 1–14, 1967.

George Spache. A new readability formula for primary-grade reading materials. *The Elementary School Journal*, 53(7):410–413, 1953.

Ulrich Tränkle and Harald Bailer. Kreuzvalidierung und neuberechnung von lesbarkeitsformeln für die deutsche sprache. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 1984.

Xuan Vinh Vo. Foreign investors and stock price crash risk: Evidence from vietnam. *International Review of Finance*, 2019.

Lester R Wheeler and Edwin H Smith. A practical readability formula for the classroom teacher in the primary grades. *Elementary English*, 31(7):397–399, 1954.

Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

Cheng Xiang, Fengwen Chen, and Qian Wang. Institutional investor inattention and stock price crash risk. *Finance Research Letters*, 2019.

Haifeng You and Xiao-jun Zhang. Financial reporting complexity and investor underreaction to 10-k information. *Review of Accounting studies*, 14(4):559–586, 2009.

Chen-Hsiang Yu and Robert C Miller. Enhancing web page readability for non-native readers. In *Proceedings of the sIGCHI conference on human factors in computing systems*, pages 2523–2532. ACM, 2010.

Jun Zhu, Amr Ahmed, and Eric P Xing. Medlda: maximum margin supervised topic models for regression and classification. In *Proceedings of the 26th annual international conference on machine learning*, pages 1257–1264. ACM, 2009.