Quantifying and analysing complicated financial text data

Joris Verhoog

(2578981)

26 October 2021

Thesis MSc Business Analytics

Master thesis committee:
Prof. dr. G. Eiben (supervisor)
Prof. dr. B. F. Heidergott (co-reader)

Preface

This thesis is written as a graduation project for the Master Business Analytics at the Vrije Universiteit while being employed as an intern at the actuarial department of EY. The report therefore follows both an academic and a business favored perspective. During the research process, I have had weekly meetings with my first supervisor Arno Hendriksen (specialized in Data Science) and meetings every three weeks with my second supervisor David Brunsveld (specialized in IFRS 17). I would like to express my gratitude towards both supervisors as they did not only provide me with subject specific information, but gave me the opportunity to gain insights into working as a consultant. I would also like to thank my VU supervisor Dr. Eiben, who organized inspiring sessions together with fellow students to share knowledge and insights. At last, I would like to thank my VU co-reader Dr. Heidergott, who provided me with valuable information on mean recurrence times.

Abstract

The exponential growth of unstructured data has led to a huge demand for analysis tools. Textual data is more difficult to handle than numerical data, especially in financial analysis. The expressions in financial articles are not as clear as the expressions on social media platforms like Twitter. Finding these expressions of sentiment can however be of great value to financial decision making. The actuarial department of the consultancy company EY is such a body that creates business value by gaining insights on the opinions of its clients. Most of the client base of the actuarial department consists of insurance companies. Many insurance companies are currently experiencing the largest change in financial reporting standards since decades. This change is called IFRS 17 and has a great impact on insurance companies that report their financial position and results on an IFRS basis and every sort of organization concerned. This thesis focuses on developing an easy-to-use tool for the actuarial department of EY to quantify and analyse opinions on specifically IFRS 17 articles from various stakeholders across the globe in the financial sector. This tool compares the distributions of sentiment on the sentences written by different geographical or industrial groups of stakeholders. The sentiment is calculated on each sentence of each article using a 'state-of-the-art' NLP method called FinBERT, which scored significantly better than a word list model. There are nine different combinations of groups that have a significant different opinion, or distribution of sentiment, according to the Mann-Whitney test for an alpha of 0.01. For example, both the consultants and the regulators separately seem to be significantly more positive compared to the insurance industry. The tool can be used to find any significant difference in opinion in a financial context and also offers the opportunity to do a more thorough analysis on the topics they are talking about either very negatively or positively.

Contents

1 Introduction

Financial reporting plays a vital role in modern businesses as it offers a level of insight in their financial positions and results. An accurate depiction of the revenues, expenses, profits, capital, and cash flow of a business is needed to provide stakeholders with in-depth financial insights. Official accounting standards are required to be able to compare businesses, bring transparency, strengthen accountability, and contribute to economic efficiency. The set of accountancy standards that need to be followed by international and stock listed companies are the International Financial Reportings Standards (IFRS) developed by the International Accountancy Standards Board (IASB). The accounts and financial reporting of these international businesses have to be officially inspected (audit) to gain compliance with IFRS. EY is an independent consulting organization that offers consulting services to businesses related to the implementation of IFRS, and is the external auditor for some other businesses. In 2017, the IASB issued IFRS 17 Insurance Contracts, which is a new set of accounting standards specifically relevant for insurance companies and will be effective in 2023. IFRS 17 is considered a fundamental change in accounting for insurance contracts, fundamentally impacting the financial reporting processes and financial information to account for insurance (and reinsurance) contracts. Besides the huge impact on insurance companies, IFRS 17 also impacts other businesses that trade contracts with the definition of "insurance contracts", such as some banks. The actuarial department of EY is a department that consults insurance companies and will provide advise on the implementation of IFRS 17 until 2023. Effective consultancy is not only about advise but also about providing information, effective diagnosis, and permanent improve of technology as to add long term value to their clients. The actuarial department of EY therefore wants to gather all relevant information on IFRS 17. A lot of information on the IFRS 17 standard is published and publicly available. However, since the implementation phase is still on and the effective date of IFRS 17 is only as of 2023, there are no results reported yet and hence also no feedback available yet from the users of IFRS 17. The opinions about the extent to which the application of IFRS 17 reaches its goals are also unknown yet, and since IFRS 17 is a fundamental change, it may take some time for the users to truly understand the results, that over time could also change opinions. Since no clear opinions on the achieved benefits or drawbacks of applying IFRS 17 are known, it will have great value if some sort of sentimental data can be found on IFRS 17. Finding or developing a method of sentiment analysis using text mining and data science adds more business value to EY compared to doing only survey research, as a data science method is reusable for similar research projects and the sentiment can perhaps be quantified and therefore statistically measured. Therefore, the goal of this thesis will be to find statistically significant opinions on IFRS 17 by organizations concerned by deploying any type of data science.

Related literature. This thesis compares two models from existing literature to calculate sentiment. The first model is based on a *Sentiment Lexicon* made by Loughran and McDonald first introduced in 2011 [1] and last updated in 2020 [2], and successfully used by De Winter and Van Dijk [3]. The second is a state-of-the-art NLP model called FinBERT trained and tested by Araci [4]. Both models were validated using the validation set of the Financial PhraseBank from Malo et al. [5]. Eventually, the NLP model scored better and is therefore used to calculate the sentiment of each sentence in each article, resulting in a distribution that can be split into groups and compared using the Mann-Whitney test as first introduced by Nachar [6]. There have been many attempts during this research to add value by creating a topic model using common literature. One of these is Latent Dirichlet Allocation created by Blei et al. [7], successfully used by De Winter and Van Dijk [3]. The application of this model on the specific data set of this research however did not show successful results. Therefore, an attempt of creating a new topic model is made by clustering the mean recurrence times of the words as presented by Berkhout and Heidergott [8]. The clustering methods from literature performed on the mean recurrence times during this research are k-medoids [9], k-means [10], DBSCAN [11], Louvain [12], agglomerative hierarchical clustering [13], Greedy Modularity Maximization [14], all also in combination with methods for dimensionality reduction like PCA [15] and t-SNE [16]. These existing methods did not show sensible results. For this reason, a custom clustering method has been developed.

Contribution. This thesis contributes to common literature as (1) it introduces a new methodology, consisting of a combination of existing models, to quantify and statistically compare opinions found in financial texts and analyse these thoroughly, (2) it successfully applies the concept of sentiment analysis using NLP on the specific subject of IFRS 17 gaining subject specific knowledge as it finds and thoroughly analyses statistically significant different opinions on IFRS 17 between different groups geographically and industrially, (3) it compares six different PDF to text extraction algorithm on an example data set and finds that three of them have a higher success rate, (4) it shows the significant higher performance of FinBERT for sentiment classification [4] against the *Sentiment Lexicon* from Loughran and McDonald [2] for sentiment classification per sentence, (5) it finds that clustering on the mean recurrence times and mean first passage times as proposed by Berkhout and Heidergott [8] shows potential for topic modeling on singular financial articles. Besides the contribution to literature only, this thesis unlocks the value of data science to contribute to businesses such as EY as (6) it provides an easy-to-use tool that makes it possible for IFRS 17 specialists without any knowledge of Python programming to do the analyses themselves, (7) this tool is also applicable to other relevant financial topics, such as the pension agreement currently in the Netherlands, (8) it provides a second easy-to-use tool, which can be seen as a by-product of this research, that can analyse financial documents individually on their sentiment and find clusters of topics using the new method of the fifth contribution.

Organization. Section 2 explains the background of IFRS 17 and why it is relevant for EY to gain knowledge about it. Section 3 shows the data found and used throughout this article and elaborates the reasoning for choosing this exact set of data regarding IFRS 17. The section also adds new dimensions by labeling the comment letters by hand (guided by an IFRS 17 specialist from EY), making it possible to compare geographical and industrial variables. Section 4 explains the NLP model used to determine the sentiment of the data. This includes model validation and comparison with other models. Section 5 analyses the distributions of the sentimental scores given by the NLP model to each sentence of each comment letter. The section shows why all the 'neutral' sentences are removed and draws attention to potential data imbalance, which has to be taken into account when taking a look at the results. Section 6 shows the results of comparing the different geographical and industrial variables using statistical testing. It will further thoroughly analyse an example pare of sentimental distributions that seems to be statistically different. This section also shows the dashboard created using Python Dash to increase the accessibility and quality of the visualization of results. Section 7 shows an extra tool created throughout the research process that did not turn to have significant valuable results for this specific research, but can be very useful for a lot of future projects. Section 8 sums up the process, draws conclusions and subsequently discusses further research directions.

2 IFRS 17

2.1 What is IFRS 17?

As an employee, customer, investor, shareholder or any kind of stakeholder of a company, you may want to know the financial position of this company. The quality of informational insights received by the stakeholder depends on the transparency and clarity of the company's financial report, and the level of assurance. The International Accounting Standards Board (IASB) sets and changes the International Financial Reporting Standards (IFRS) with a goal of "providing a high quality, internationally recognised set of accounting standards that bring transparency, accountability and efficiency to financial markets around the world" [17]. The work of the IASB "serves the public interest by fostering trust, growth and long-term financial stability in the global economy" [18]. The number posterior of IFRS specifies a category or aspect relevant to financial reporting. For example, IFRS's relevant to many businesses in the financial sector are IFRS 9 (financial instruments), IFRS 3 (fair value measurements), and IFRS 16 Leases just to name a few. IFRS 4 is specifically used for insurance contracts and came into effect in 2005 [19]. IFRS 17 serves to replace IFRS 4, initially by January 1st 2021. The goal of IFRS 17 as stated by the IASB is: "IFRS 17 Insurance Contracts establishes principles for the recognition, measurement, presentation and disclosure of insurance contracts within the scope of the Standard. The objective of IFRS 17 is to ensure that an entity provides relevant information that faithfully represents those contracts. This information gives a basis for users of financial statements to assess the effect that insurance contracts have on the entity's financial position, financial performance and cash flows" [20]. It provides a robust standard as IFRS 4 allows a multitude of different accounting policies even within insurance, resulting a lack of comparability between insurance businesses and between the insurance sector and other sectors. IFRS 17 includes complex fundamental changes to accounting in liability measurement and profitability recognition for insurance contracts. The main changes in IFRS 17 concern the measurements of the insurance contracts liabilities, and a contractual service margin. There are three measurement models for different contract types: the Building Block Approach (BBA), the Premium Allocation Approach (PAA), and the Variable Fee Approach (VFA) [21].

BBA is the default model for the measurement of insurance contract liabilities. One key and new component in the insurance contract liabilities as measured by the BBA and VFA approaches is the Contractual Service Margin (CSM) [22]. The CSM represents the expected remaining profit on the insurance contract. See Figure 1 for an example of calculating the CSM at initial recognition of an insurance contract.

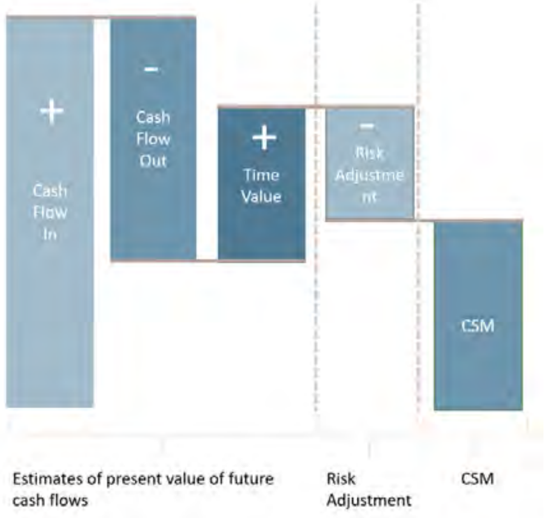


Figure 1: Calculation of CSM. Example: *Suppose there is a 15 year insurance contract with an upfront premium of 90 euros and an expected claims payout of 100 euros in the 10th year. At the inception of the contract, the sum of the future cash flows is 90 euros minus 100 euros is* $90 - 100 = -10$ *euros. The present value of the cash inflow is 90 euros and the present value of the 100 euros outflow at a 3% flat discount rate is* $100 * 0.97^{10} = 74.74$ *euros. So taking into account the time value of cash flows, the liability is now -16.26 euros. With a risk adjustment for the uncertainty in the amount and timing of the claims of 5 euros, the net liability is* $-16.26 + 5 = -11.26$ *euros. 11.26 euros is the CSM, which is the expected remaining profit of the contract.*

Now, we have the CSM. During the coverage period, the total liability has two components. The first component is the 'liability for remaining coverage' that exists because cash flows at the beginning of the period has not been earned. The insurer has not been released from risk yet, therefore the cash flow its time value and its risk adjustment are considered as liability. The second component is the 'liability for incurred claims' that refers to unpaid claims for insured events that have happened. Insurers are not always aware of every incurred claim at the current accounting period. Therefore the liability is calculated as a risk adjusted discounted cash flow. The building blocks are treated differently during profit recognition. Changes in cash flows and risk adjustment relating to future service adjust the CSM, which is amortized to Profit Loss (P&L) over time, while those relating to past and current services flow into P&L.

An alternative simpler model called PAA can be used for some short term contracts (<1 year) in calculating the pre-claims liability. It is a much simpler model than BBA. Premiums received on day 1, or unearned premium reserve, along with any upfront acquisition cost component are recorded as a current liability and as cash in the asset. This is because the coverage period is yet to pass and the premium has not actually been earned. As insurers are gradually released from risk as time goes by, the unearned premium flows from liability to revenue account.

The VFA is a variation of the BBA. It applies to contracts with direct participation features, such as for many Unit-Linked type of insurance contracts. These contracts are life and insured pension contracts. Investor's fund is invested into a mixture of assets to average out the volatility in the market, smoothing out returns. A substantial share of returns from these items are paid out. For these contracts, the VFA is most suitable as the contractual cash flows also depend on the returns on underlying items (such as specified assets). The principle of the VFA is similar to BBA, but the changes in assets supporting insurer's share is adjusting the CSM, whereas for the BBA the effect of such changes would not adjust the CSM. The CSM is accreted at current interest rate, and changes in the value of options and guarantees are recognised in CSM but are allowed present in P&L when there is risk mitigation.

IFRS 17 is a technology-heavy change program, as there is a large increase in the amount of data points needed to account for insurance contracts compared to today. These data points will also be sourced from multiple systems, such as base administration systems, payments systems, actuarial model environment (projection systems), and accounting systems. Its impact would be profound not only on financial and actuarial function but also on other business functions such as product design and business planning. It needs further requirements in data systems on top of existing frameworks, such as Solvency II. Calculation of CSM is a major undertaking for data collection, storage and processing, and IT architecture as it requires much more granular levels of measurement. IT, finance, risk, actuarial, and business teams are all internal stakeholders. IFRS 17 introduces greater volatility on P&L which means it requires more advanced forecasting and simulation capabilities to make financial forecasts, which is likely also relevant to external stakeholders, such as investors and sector analysts.

In summary, IFRS 17 is a financial reporting standard for insurance companies that has to be implemented before 2023 for insurers reporting on IFRS, and is a fundamental change to the insurance sector globally. It specifies which liability calculation models (BBA, PAA & VFA) to use and how it is used in each type of each insurance contract. It is a major technological undertaking and has wide impact on the operating process of other functions as well as data administration, financial presentation and actuarial calculations will need to change.

After the IASB issued IFRS 17 on May 16th 2017, the IASB proposed amendments to IFRS 17 on June 26th 2019 to narrow-scope the IFRS standards (see Figure 2). The fundamental principles introduced in May 2017 by the IASB remained unaffected. The amendments were aimed at helping companies implement the Standard and making it easier for them to explain their financial performance. Together with the announcement of the amendments, the IASB eased transition by deferring the effective date of the Standard from 2021 to 2023.

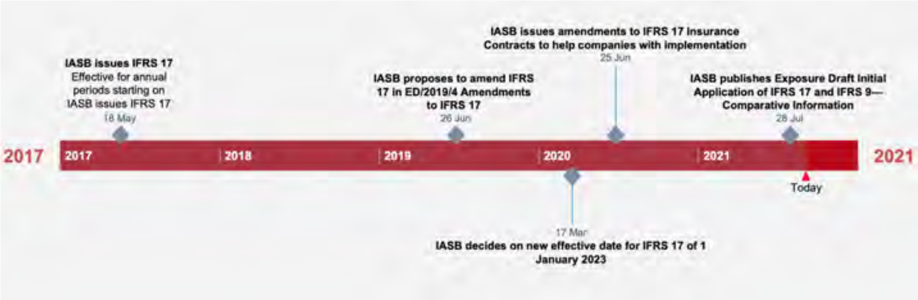

IASB issues amendments to IFRS 17 Insurance Contracts to help companies with implementation
25 Jun
IASB issues IFRS 17
Effective for annual periods starting on IASB issues IFRS 17
18 May
IASB proposes to amend IFRS 17 in ED/2019/4 Amendments to IFRS 17
26 Jun
IASB publishes Exposure Draft Initial Application of IFRS 17 and IFRS 9—Comparative Information
28 Jul
2017
2017
2018
2019
2020
2021
2021
Today
17 Mar
IASB decides on new effective date for IFRS 17 of 1 January 2023

Figure 2: The overall timeline of IFRS 17 from the moment it was issued until now.

2.2 Why is IFRS 17 relevant for EY?

EY is a consulting company eager to help third companies gain value. The actuarial department of EY mainly consults insurance companies. Example activities of EY actuarial are giving support or advice in calculating portfolio risks, managing pension funds, interpreting and implementing accountancy frameworks and reviewing various valuations underlying to financial reports. Since a few years, IFRS 17 has become a leading part of their activities. Implementing IFRS 17 poses many challenges to insurers, such as how to interpret the standard and adjust methodologies, systems, and processes accordingly. There are no financial reports yet published that apply IFRS 17, and insurers need to implement IFRS 17 next to their existing processes. Consultants can support insurers with the activities needed or address challenges, and leverage their knowledge build up from their (large and global) networks and insights to how other insurers are doing it. Not until the effective date of January 1st 2023, any official practice data can be released that will show the potential drawbacks or benefits of IFRS 17. The only data available right now are expectations and opinions on IFRS 17 from organizations concerned.

Being able to quantify these opinions on IFRS 17 from the organizations concerned would help EY gain more insights in the organizations, and their concerns, outlook, and market trends. With this information, EY can give more accurate advice and add value using deeper, broader (and scaleable) insights such as from opinions of specific organizations. If some organization is for example very negative about a specific subject regarding IFRS 17 relative to similar organisations, or other organisations in the same sector, EY can then approach this organization by explaining more on the benefits and drawbacks of this specific subject relevant to this organization. Also, when we see that a certain industry is more positive than a different industry, EY gains knowledge on the way of thinking of organizations in this industry. This is especially valuable when clients are within this industry. More knowledge on the way of thinking of your client means more valuable and goal oriented consult. Further, being able to quantify the opinions from many stakeholders using public available data, allows EY to compare their own opinions against opinions that are independently quantified, and in that case identify the specific subjects of IFRS 17 underlying to those quantified assumptions.

2.3 Research questions

The research question that we try to answer is:

- "Can we find statistically significant opinions on IFRS 17 from the organizations concerned?"

If this question gets affirmed, the follow-up questions are:

- "Can we find the topics that organizations are talking about when they are either very negative or positive?"
- "Can we find some of the most negative and positive comments on IFRS 17?"

3 The data

3.1 Finding the data set

It is of great importance to find a suitable data set that can help give answers to the research questions. Finding such a data set is however challenging as there is no easily accessible forum like Twitter where financial organizations react on standards like IFRS 17. The challenge of finding a suitable data set has been an extremely time consuming part of the process. First attempts of finding suitable articles are done by searching IFRS 17 on financial websites that have a high probability of containing articles with an opinion. Examples of organizations that publish financial news articles with an opinion are credit rating agencies like Moody's Investors Service, Fitch Rating, and Standards Poor's. The number of articles published by these on IFRS 17 are however very limited and the articles that are published are sometimes only a few sentences long making a short statement. When there seems to be availability of a large document, the organization requests a payed membership before being able to open and read the documents. Also after a more broadly search for financial news on IFRS 17 on a search machine like Google, the number of news articles that pop up are again either limited and inaccessible. The only available articles are just a few and not even written on similar events of the IFRS 17 timeline (Figure 2). To be able to do a comparison on the opinions of different organizations, they have to be talking about the same event though. Searching for data sets and testing them with a multitude of data science models like NLP did not result in any usable outcome for this research until this moment, where a few months passed. Therefore, a couple of meetings have been setup with EY's IFRS 17 specialists to gain information and ask for tips. After several meetings, the golden tip came from a global leader and happened to be the search phrase 'comment letters'. Comment letters are letters written by organizations concerned that react on a certain event. With that in mind, a search has been done on comment letters per event on the IFRS 17 timeline (Figure 2). Every time that the IASB issued a new exposure draft, the IASB also asked to public (organizations concerned) to react on this via a comment letter. The exposure draft that got the most comment letters is the one where the IASB proposed to amend IFRS 17 in ED/2019/4 Amendments to IFRS 17. To be precise, there are comment letters from 123 different organization from all over the world and from many different industries. Fortunately, the IASB gathered all the comment letters on this exposure draft and posted them publicly on the IASB website. This data set has the highest chance to create the opportunity to discover the opinions of these organization and perhaps the difference between any geographical or industrial factors.

3.2 Labeling the data set

The comment letters however do not specifically specify the industry and the operational country or head office of the organization that has written it. Being able to group the comment letters on geographical or industrial features creates the possibility to perform statistical research on potential differences between these groups. To be able to group the different comment letters, these comment letters have to be labeled by hand. This labeling is done in cooperation with a IFRS 17 specialist of EY. The labels given to each organization that commented on ED/2019/4 Amendments to IFRS 17 are the continent of operation (op_continent), type of organization (org_type), industry type (industry), and, if the industry type equals 'insurance', then also the insurance type (insurance). Since most clients of the actuarial department of EY are from the insurance industry, it is relevant to search for potential differences in opinions between the different types of insurance: life, non-life, health, reinsurance, and composite. The different labels for each variable and the number of organization that got this label is shown for each label of each variable in Table 1. The reason that 'health' is not a label for the variable 'insur_branch' is the fact that there are no comment letters written by just a health insurance organization. Initially, also a fifth variable was added, which is the country of the head office (head_office) of the organization that commented. However, this label was not used during further analysis and research. This since there are too many different countries of head office meaning that the count of organizations that correspond to each label of country of head office is too low to do any analysis or comparison research. The initial table including the country of the head office can be found in Table 10 in the Appendix in Section 10.1.

op_continent	Nr.	org_type	Nr.	industry	Nr.	insur_branch	Nr.
Africa	7	Academic	2	Academic	2	Composite	22
Asia	27	Company	38	Accounting	17	Life	12
Europe	37	Regulation committee	24	Actuarial	13	Non-life	5
North-America	9	Sector committee	59	Banking	11	Reinsurance	3
Oceania	7			Consulting	10		
South-America	3			Credit rating	2		
Worldwide	33			Insurance	42		
				Regulation	24		

Table 1: The labels given to each comment letter and the corresponding number of documents that got this label.

3.3 Extracting the data

The comment letters have to be extracted to text data before being able to analyse them. All comment letters are in PDF form and the programming is done in Python. Therefore the most effective PDF to text extracting package in Python has to be found. Unfortunately, no package was found that could extract the text of all the files successfully. Unsuccessful extraction means that either an error was given when trying to open the PDF files or the extracted text happened to turn out empty. Several packages were tested to find the one that works most effective on this specific data set. The number of PDF files that are successfully extracted are shown in Table 2 per package.

Package	Nr. of successfully extracted files	Nr. of different writers
PyPDF2	88	85
pdfminer	89	86
pdfplumber	95	92
tika	104	98
pdftotext	104	98
fitz	104	98

Table 2: The labels given to each comment letter.

The three packages with the highest ratio of successful text extractions are *tika*, *pdftotext* and *fitz*. Eventually, the conclusion was made that *fitz* was the best choice for the continuation of this research. This is due to the fact that *fitz* already filtered out the many enters (’\n’) and tabs (’\t’) the documents contained, which *tika* and *pdftotext* did not.

So during extraction of the PDF documents, it was inevitable that a part of the data got lost. From the total of 123 organization that sent in a comment letters, the 98 organizations left over after extraction into Python are shown in Table 3 with the count per label.

op_continent	Nr.	org_type	Nr.	industry	Nr.	insur_branch	Nr.
Africa	~~7~~→5	Academic	2	Academic	2	Composite	~~22~~→13
Asia	~~27~~→24	Company	~~38~~→26	Accounting	~~17~~→15	Life	~~12~~→9
Europe	~~37~~→27	Regulation committee	~~24~~→18	Actuarial	13	Non-life	~~5~~→4
North-America	9	Sector committee	~~59~~→51	Banking	~~11~~→8	Reinsurance	~~3~~→2
Oceania	~~7~~→6			Consulting	~~10~~→9		
South-America	3			Credit rating	2		
Worldwide	~~33~~→23			Insurance	~~42~~→28		
				Regulation	~~24~~→18		

Table 3: The labels given to each comment letter and the corresponding number of documents that got this label and got successfully extracted to Python.

4 Sentiment classification models

After extracting the PDF files into text using the *fitz* Python package, a method can be used to calculate the sentiment in each article. Two separate methods are used to calculate the sentiment. The first method of calculating sentiment is using a pre-defined word list, and the second method is using a state-of-the-art machine learning model called FinBERT.

4.1 Word list model

The first method is based on a (not yet published) article written by De Winter and Van Dijk [3]. They calculate the sentiment on 1 million news articles published by *Financieel Dagblad* from 1985 until now. To calculate the sentiment in each article, they use the *Sentiment Lexicon* of Loughran and McDonald first introduced in 2011 [1] and last updated in 2020 [2]. This is a known list of sentimental words commonly used in financial articles, both negative and positive. De Winter and Van Dijk translated the English word list to Dutch since the articles of Financieel Dagblad are also written in Dutch. By counting the number of times these words appear in an article, -1 for a negative word and +1 for a positive word, they calculate the sentiment per article. Using a moving average window, the calculated sentiment on these financial articles successfully follows the Dutch GDP over time as can be seen in Figure 3.

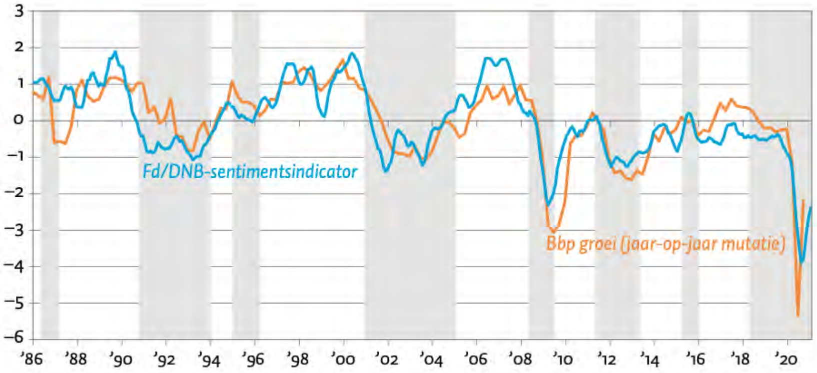


Figure 3: The Dutch GDP from '85 until '20 (in orange) against the moving average of the calculated sentiment on the articles of *Financieel Dagblad* (in blue). The calculated sentiment follows the Dutch GDP closely. [3]

In their article, De Winter and Van Dijk calculate the sentiment per article. However, in this research case we are going to calculate the sentiment per sentence. This is due to the fact that our validation set exists of sentiment labeled sentences.

4.2 Natural Language Processing model

The second method of classifying sentiment is less transparent. This classification is done using a machine learning model called Bidirectional Encoder Representations from Transformers, or in short: BERT. BERT is a natural language processing model that makes use of transformers, which are a fairly new family of neural networks architectures. A transformer is a deep learning model that weights the significance of each part of the input data by remembering the relevant parts and forgetting irrelevant parts. In the field of language modelling, they are the state-of-the-art family of neural network architectures especially compared to more common families of neural networks architectures like CNN's [23] and RNN's [24].

4.2.1 Transformers

Convolutional Neural Networks (CNN) work by moving a sliding window over an input array and projecting that into a smaller array. These convolutions are then stacked on top of each other. In general, these types of models work best for image data, and have not had quite as much success with natural language data. Recurrent Neural Networks (RNN) work by moving over a sequence. At each point in time, they look at both the information available from the current point of sequence and from sequence points before that one. They have had a lot of success with natural language data, but they do have some drawbacks. One of the biggest drawbacks is that you need to look at the sequence in order, which means there is a limit to how much you can parallelize the training on different machines. This results in a bottleneck in how fast you can train these models. Another big drawback is that it is easier for RNN's to capture relationships between points that are close to each other compared to capturing relationships between points that are very far from each other, say several thousands points in a sequence. Transformers were proposed to help address these problems. What sets transformers apart from other architectures is *self attention*, as the title of the paper implies: *"Attention Is All You Need"* [25]. The general idea of attention is that you learn a weight between each input item and each output item. Self attention is very similar, but instead of looking for a relationship between the input and the output, it looks at the relationship between each item in the input sequence and every other item in the input sequence. Multi-headed self attention does this several times, each learning a different focus of attention. The original transformer architecture uses a traditional sequence to sequence model that contains an encoder that changes the input sequence into embeddings, which are numeric representations of that input sequence, and a decoder that turns the embeddings into the desired output. Both the encoder and the decoder are made up of a number of multi-headed self attention modules that are stacked on top of each other (see Figure 4). Transformers were originally proposed for sequence to sequence modelling and specifically machine translation. However, BERT uses the architecture of the transformer a bit different.

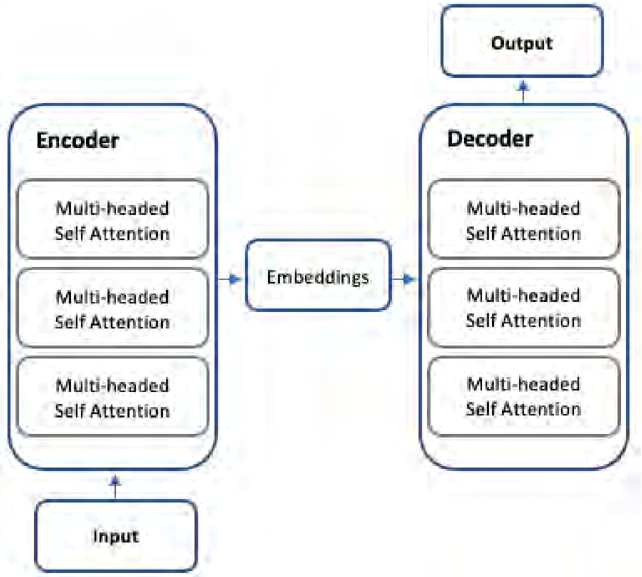

Output
Encoder
Decoder
Multi-headed Self Attention
Multi-headed Self Attention
Multi-headed Self Attention
Embeddings
Multi-headed Self Attention
Multi-headed Self Attention
Multi-headed Self Attention
Input

Figure 4: The architecture of a transformer.

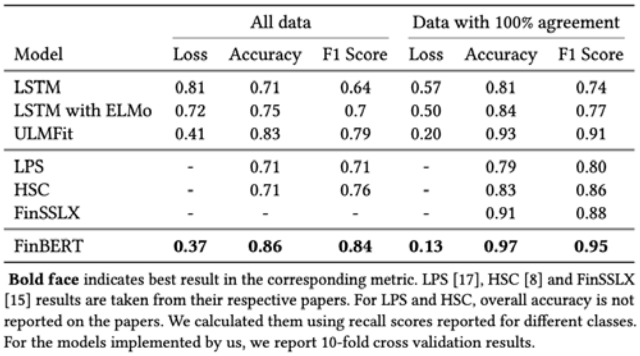
4.2.2 BERT

BERT takes multiple transformer encoders, either 12 (BERT-Base) or 24 (BERT-Large), and stacks them on top of each other. It then trains these encoders by removing words from the input sequences and having the model fill in the blank spaces, which makes BERT a masked language model. This way, BERT is bidirectionally trained, which means that it takes both the previous and next tokens into account at the same time. Masked language models are useful as they are one way of creating *contextual* word embeddings, meaning different embeddings are created for the same words having a different meaning. For example, ”lie” can either mean someone ”lying down” or someone ”being untruthful” depending on the context. This masked learning is called the pre-training of BERT. Pre-training can take months with a single GPU, meaning it will be impossible to train BERT ourselves within the time frame of this research. However, since the paper of BERT was published there has been done a lot of pre-trained work on more than 100 different languages all openly available on the internet.

One way to use BERT after pre-training is to use the embeddings, present in the last layer of the network, as an input for a different classifier. A more common way is to fine-tune the original pre-trained model towards a specific task. A small output layer will be added to the original BERT model and the weights will then actually be updated without substantial task-specific architecture modifications. Fine-tuning towards a specific task requires much less data since most of the training has already been done during pre-training. Example tasks that BERT can be fine-tuned on are question answering, text summarizing, named entity recognition, sentiment classification, and more. [26]

4.2.3 FinBERT for financial sentiment analysis

In the scope of this research, the desired task of the BERT model is to classify financial texts on sentiment. The specific pre-trained model that suits this case is FinBERT for financial sentiment analysis, which is trained and tested by Araci [4]. The model is first pre-trained to learn the english language on an english Wikipedia corpus and a books corpus, which in total contain more than 3.5 billion words. Then the model is further pre-trained to specifically learn the financial english language using TRC2-financial [27], which is a data set of financial news articles that contain approximately 29 million words. After these two steps of unsupervised training follows the fine-tuning part that pushes the model to the specific task of sentiment classification. The data set used is Financial PhraseBank from Malo et al. [5], which consists of 4845 sentences selected randomly from financial news articles that are labeled by 16 annotators with background in finance and business. The data set also includes information regarding the agreement levels on sentences among annotators. 20% of the total sentences were set aside as test test and 10% of the remaining sentences were set aside as validation set, which eventually leaves 3488 sentences available for training. As can be seen in the paper, the model clearly outperformed all baseline methods implemented by themselves, such as ULMFit [28] and LSTM [29] with ELMo embeddings [30], and also models reported by other papers.



	All data			Data with 100% agreement		
Model	Loss	Accuracy	F1 Score	Loss	Accuracy	F1 Score
LSTM	0.81	0.71	0.64	0.57	0.81	0.74
LSTM with ELMo	0.72	0.75	0.7	0.50	0.84	0.77
ULMFit	0.41	0.83	0.79	0.20	0.93	0.91
LPS	-	0.71	0.71	-	0.79	0.80
HSC	-	0.71	0.76	-	0.83	0.86
FinSSLX	-	-	-	-	0.91	0.88
FinBERT	**0.37**	**0.86**	**0.84**	**0.13**	**0.97**	**0.95**

Bold face indicates best result in the corresponding metric. LPS [17], HSC [8] and FinSSLX [15] results are taken from their respective papers. For LPS and HSC, overall accuracy is not reported on the papers. We calculated them using recall scores reported for different classes. For the models implemented by us, we report 10-fold cross validation results.

Figure 5: Results of FinBERT compared to other NLP methods. [4]

4.3 Models validation

To recap, the word list model does not have to be trained and it is more transparent since it is exactly known on which words the model classifies. This is unknown for the FinBERT model, which gives a sentiment score based on millions of different parameters. To compare the Loughran and McDonald word list model and the FinBERT model for financial sentiment analysis on their performance, both models are validated using the remaining unused validation set of the Financial PhraseBank data set. This validation set is chosen since sentiment labeled financial text data is very rare and this data set is suitable to validate both models. The validation set contains 388 sentences, of which 182 sentences have an agreement level of 100%. The accuracy's of both models on the complete validation set and on the validation set with an agreement level of 100% only are shown in Table 4. It shows that FinBERT performs significantly better compared to the word list method. This method will therefore be used to calculate the sentiment scores for each sentence of each comment letters for the continuation of this project.

	Accuracy on validation set with:	
Method	**agreement level >50%**	**agreement level =100%**
Word list	0.58	0.65
FinBERT	0.85	0.97

Table 4: The accuracy score per method of sentiment calculation.

5 Sentiment data

5.1 Distribution of the sentimental data

FinBERT is used to calculate the sentiment per sentence. The input of the model is a string sentence and the output is an array containing three numbers between 0 and 1. These numbers represent the probabilities of the sentence being positive, negative, and neutral, in that specific order. The label with the highest probability will be the *prediction* given to the sentence. The *sentiment score* is calculated by subtracting the probability of the sentence being negative from the probability of the sentence being positive. This means that the closer the *sentiment score* to -1, the more negative and the closer the *sentiment score* to +1, the more positive the sentence. The *sentiment scores* will be the data points used to measure sentiment from now on. Since the number of sentiment scores is equal to the number of total sentences of all the comment letters, this results in a lot of data points that can be plotted in a single distribution as shown in Figure 6.

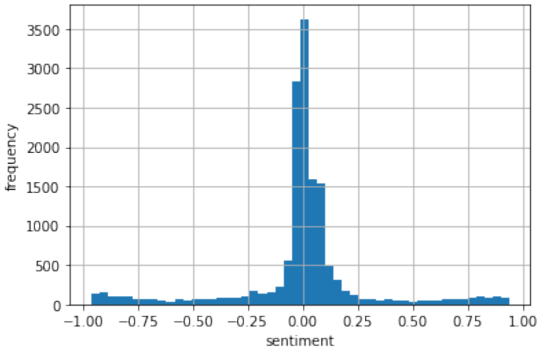
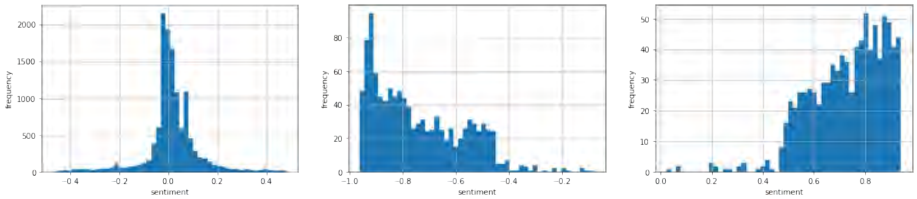


Figure 6: The distribution of the sentiment scores on every sentences of each comment letter.

To compare the sentiment of different groups with each other, the data can be filtered on these two groups only (for example Europe and Asia) and a statistical test can be performed on whether these come from the same distribution. A well-known statistical test that does this is the two sample t-test [31]. The t-test however assumes that both distributions are sampled from a normal distribution. Therefore, before being able to use this test, it is necessary to check if the data is normally distributed. The distribution of the sentiment scores of all sentences however does not look like it is normally distributed as the tails are thicker at the complete ends and the peak in the middle is extremely high. The thick tails on the left and right consist of respectively the negative and positive labeled sentences, and the high peak in the middle consists of the neutral labeled sentences. When separating the data into the three different labels of prediction (*positive*, *negative*, and *neutral*), the distributions perhaps come closer to a normal distribution. These separate distributions are shown in Figure 7.



(a) Neutral (b) Negative (c) Positive

Figure 7: The distribution of the sentiment scores on every sentences of each comment letter divided into the three prediction labels *neutral*, *negative* and *positive*.

As can be seen in Figure 7a, the distribution of the *neutral* labeled sentences seems to come a little closer to the normal distribution, but not enough. The thick tails of the complete distribution of sentiment scores are gone, but it still has a very high peak (high kurtosis [32]). As can be seen in Figures 7b and 7c, the separate distributions of the *negative* and *positive* labeled sentences both visually do not seem to be normally distributed. The distribution of the *negative* labeled sentiment scores and the *positive* labeled sentiments respectively seem to be extremely left skewed and right skewed. The expectation that all distributions do not seem normally distributed gets confirmed when we statistically test all distributions on the null hypothesis: 'the data comes from a normal distribution' with the Kolmogorov-Smirnov test [33], the Shapriro-Wilk test [34], and the D'Agostino-Pearson test [35]. As can be seen in Table 5, all p-values are far below 0.01 meaning that no distribution is even close to being normally distributed.

		All	**Neutral**	**Negative**	**Positive**
Kolmogorov-Smirnov test	**Statistic**	0.654	0.383	0.637	0.654
	p-value	0.0	0.0	0.0	0.0
Shapiro-Wilk test	**Statistic**	0.938	0.870	0.920	0.938
	p-value	7.43e-19	0.0	1.19e-23	7.43e-19
D'Agostino-Pearson test	**Statistic**	118.9	1438.6	87.13	118.9
	p-value	1.52e-26	0.0	1.20e-19	1.52e-26

Table 5: The test statistic and the corresponding p-value calculated using the Kolmogorov-Smirnov test, the Shapriro-Wilk test, and the D'Agostino-Pearson test on the complete distribution of sentences, the neutral labeled sentences, the negative labeled sentences, and the positive labeled sentences.

The chance that a distribution filtered on a single group is normally distributed is even smaller since the frequency of data will be lower. As the full distributions of data already get strongly rejected to originate from a normal distribution, this means that the t-test definitely can not be performed to compare different distributions to answer questions on factorial differences. Instead, a non-parametric test called the Mann-Whitney test will be used, which does not assume normality when comparing. The null hypothesis for the Mann-Whitney test is: "the two samples come from the same distribution". The Mann-Whitney test both looks at the difference in mean and median of the two different distributions. The drawback of this test is that the findings are less strong compared to the t-test. However, the benefit is that we do not have to prove for every distribution that it is normally distributed.

5.2 Cleaning the data

During analysis of the prediction labels given to the sentences, two issues were encountered on the neutral labeled sentences. Firstly, there are way more neutral sentences than sentimental sentences as can be seen in the box plots in Figure 8. This means that finding a significant difference between two distributions according to the Mann-Whitney test is possibly due to the neutral sentences of one group having a different sentiment score than the neutral sentences of the other group. Secondly, just by scanning the individual comment letters by hand, it strikes that some comment letters copy and paste large paragraphs written by the IASB in the exposure draft before they comment on it. Almost all of these sentences do not contain any sentiment as they are all informative. Other comment letters do not repeat these paragraphs and comment immediately. This means that the data is possibly disproportional, where some comment letters have much more neutral sentences than the other. As can be seen in the box plot in Figure 8, the percentage of neutral sentences can differ between less than 50% and up to 100%. Now in theory, two groups can have the exact same opinion, but the Mann-Whitney test can still witness a significant difference in distributions if one group copies sentences from the Exposure Draft written by the IASB and the other group does not.

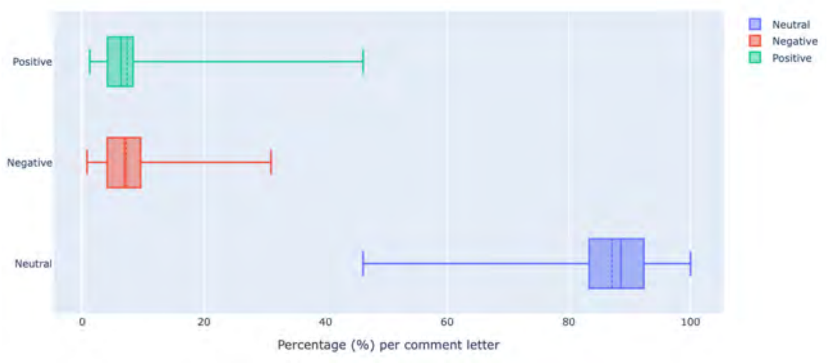


Figure 8: Box plots on the distribution of the percentages of the sentences per comment letter containing each prediction label (*positive*, *negative*, and *neutral*) separately.

Due to these two main reasons, the choice has been made to remove all neutral labeled sentences from the data set. As we do not have to assume any normality in the data, it is possible to compare the distributions with these sentences removed. The distribution that is left will then look like the one in Figure 9, where the *negative* and *positive* labeled sentiment scores are each shown in a different color. There are some positive and negative labeled sentences with a sentiment score close to zero. This is due to the fact that the sentiment score is calculated by subtracting the probability of the sentence being negative from the probability of the sentence being positive. If these probabilities only differ a small amount, the sentiment score is also very small. The reason for the sentence not being labeled neutral in this case, is the fact that the probability of the sentence being neutral is even lower.

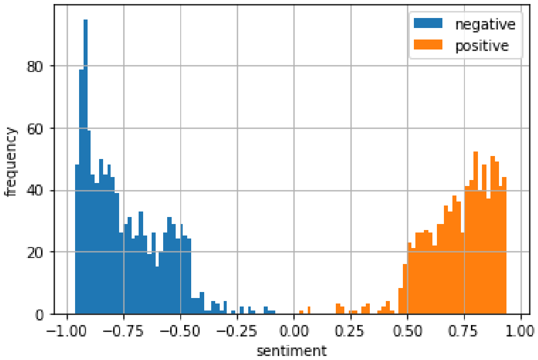


Figure 9: The distribution of all the sentences of each comment letter labeled with sentiment. The *negative* labeled sentences are shown in blue and the *positive* labeled sentences are shown in orange.

The distribution shown in Figure 9 will be used to do factorial comparisons within the data set. When the Mann-Whitney test proves a significant difference between two groups, it means that one distribution tends significantly more towards the left (negative) or towards the right (positive). When this happens, the conclusion can be made that one distribution of sentences is more positive or negative compared to the other. In other words, it is then possible to state that one group has a statistically significantly different opinion than another group.

5.3 Imbalance in the data

After labeling, classifying, and cleaning the data, it is important to check any imbalance in the data. When finding a significant difference in sentiment between two groups, this conclusion is possibly due to an indirect consequence of data imbalance. To explain this, let's use an example: as can be seen in Figure 10, the data from Africa mostly comes from the accounting industry and the data from South-America mostly come from regulators. So if one of our conclusions is: "Africa is more negative than South-America", then it is important to check whether accountancy is also significantly more negative than regulators. If that is true, then one of the conclusions is an indirect consequence of the conclusion of the other, making one of the conclusions possibly invalid. Analysing any imbalance in the data will be part of the evaluation process of the results in the following section.

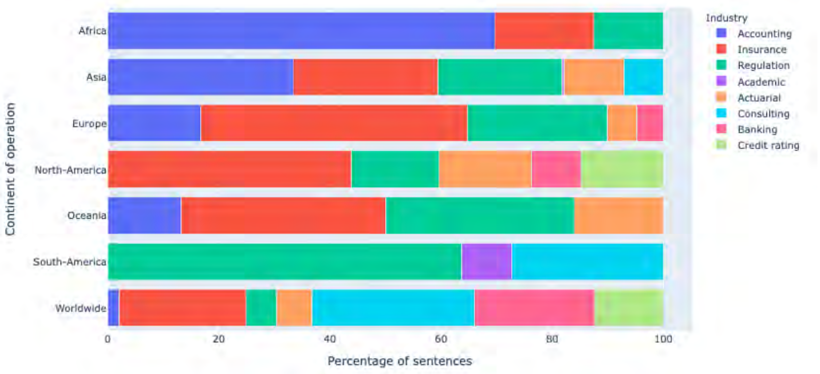


Figure 10: The distribution (%) of the industries per continent of operation.

6 Results

6.1 Statistical testing

To answer the research question ”Can we find any statistically different opinions on IFRS 17 from the organizations concerned?”, we check any significant differences in sentiment between different geographical and industrial groups using the non-parametric Mann-Whitney test since the sentiment distribution is not normally distributed (elaborated in Section 5.1). First, let’s have a look at the distributions of the values for the industry variable. Figure 11 shows the box plots of the distributions of the data per industry. As can be seen, the consulting industry has the highest sentiment and the insurance industry has the lowest sentiment regarding both the mean (dotted line) and the median (unbroken line).

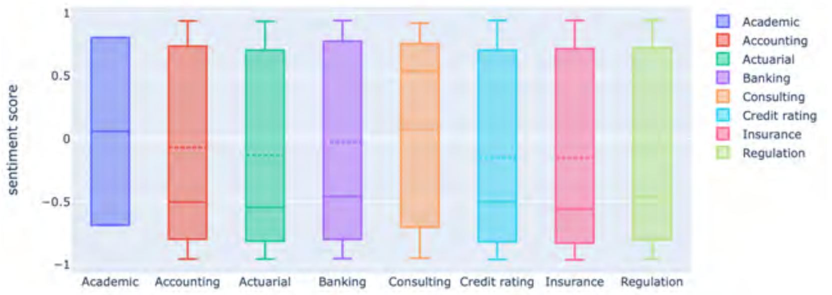


Figure 11: Box plots of the distributions of sentiment scores for each industry separately.

The p-value of the Mann-Whitney test is calculated for every possible combination of two industries, which is shown in table 6. Again, the null hypothesis for the Mann-Whitney test is: ”the two samples come from the same distribution”. The p-values between 0.05 and 0.01 are colored in orange and the p-values below 0.01 are colored in red.

	Academic	**Accounting**	**Actuarial**	**Banking**	**Consulting**	**Credit rating**	**Insurance**
Accounting	0.68254						
Actuarial	0.57509	0.47145					
Banking	0.67194	0.43977	0.2116				
Consulting	0.85775	0.12186	0.03925	0.0.5651			
Credit rating	0.59512	0.481	0.90471	0.26695	0.06356		
Insurance	0.55528	0.12961	0.63528	0.05198	0.00415	0.83656	
Regulation	0.66932	0.86947	0.34994	0.47013	0.10775	0.38967	0.0476

Table 6: The p-values of the Mann-Whitney test performed on every combination of two industries.

As can be seen in table 6, the only combination of two industries that significantly rejects the null hypothesis with a p-value below 0.01 is Insurance-Consulting. This means that we can statistically state that the insurance industry has a different opinion compared to the consulting industry. The large difference in median and mean was also noticeable in the box plots shown in Figure 11, which shows that consulting is the most positive and insurance the most negative.

A different variable that shows statistically significant different opinions between groups is the continent of operation. The p-value of the Mann-Whitney test is calculated for every possible combination of two continents of operation and shown in Table 7, where the p-values between 0.05 and 0.01 are colored in orange and the p-values below 0.01 are colored in red.

	Africa	**Asia**	**Oceania**	**Europe**	**North-America**	**South-America**
Asia	0.44572					
Oceania	0.34978	0.00924				
Europe	0.75588	0.26525	0.04113			
North-America	0.8673	0.076	0.33537	0.25694		
South-America	0.3221	0.49013	0.0693	0.32979	0.23778	
Worldwide	0.406	0.92599	0.00587	0.15984	0.04854	0.54692

Table 7: The p-values of the Mann-Whitney test performed on every combination of two continents of operation.

When taking a look at the p-values shown in Table 7, the null hypothesis is definitely rejected for the combinations of Oceania-Asia and Worldwide-Oceania as these p-values are below 0.01. Therefore, the conclusion can be statistically made that the comment letters from Oceania show a different opinion compared to the comment letters from Asia or Worldwide.

The comparison between the distribution of groups can also be more specific by applying a filter. For example, instead of comparing all organizations between different continents of operation as done in Table 7, it is also possible to compare only the companies between different continents of operation. This is done in Table 8, which shows the calculated p-values using the Mann-Whitney test on all possible combinations of continents of operations filtered on companies only.

	Africa	**Asia**	**Europe**	**North-America**	**South-America**
Asia	0.44041				
Europe	0.99412	0.0364			
North-America	0.80786	0.14091	0.47692		
South-America	0.93265	0.69784	0.91025	0.97023	
Worldwide	0.32021	0.54528	0.00216	0.01343	0.63304

Table 8: The p-values of the Mann-Whitney test performed on every combination of two continents of operation (filtered on companies only).

Table 8 shows presence of a significant difference in sentiment between the companies operating in Europe versus the companies operating Worldwide, since the calculated p-value between those two is smaller than 0.01.

Using the previous three tables, it was already possible to state four statistically significant differences in opinions on IFRS 17 between different groups geographically and industrially (using the Mann-Whitney test and an alpha of 0.01). The performed comparisons of opinions in these three tables are just a few of many more comparisons possible. These three tables are also shown in the Appendix (Section 10) together with all the other relevant tables containing comparisons using the p-values of the Mann-Whitney test. The tables in this part of the Appendix show the comparisons of continents of operation (Section 10.2) and industries (Section 10.3) as shown above, but also with filters on each possible type of organization separately (Section 10.2-10.2 and 10.3-10.3). The three types of organization are compared as well (Section 10.4), both with no filter and with a filter on each of the continents of operation separately (Section 10.4-10.4). Also, the insurance types are compared with each other (Section 10.3), both with no filter and with a filter on each type of organization separately (Section 10.5-10.5), except for the type of organization 'regulator' since there are no insurance regulators.

Out of all the comparisons in the Appendix, the ones that show a statistically significant difference between the two distribution for a p-value lower than 0.01 are shown in Table 9. There are a lot of distributions that also show a slightly less significant difference, namely for a p-value lower than 0.05. These are shown in the Appendix, Section 10.6 in Table 29.

Mann-Whitney test input			
The two distributions		**Filtered on**	**p-value**
Oceania(-)	**Worldwide(+)**	None	0.00065
Europe(-)	**Worldwide(+)**	Company	0.00216
Insurance(-)	**Regulation(+)**	None	0.00348
Insurance(-)	**Consulting(+)**	None	0.00415
Company(-)	**Regulation committee(+)**	Europe	0.00580
North-America(-)	**Worldwide(+)**	None	0.00604
Oceania(-)	**Africa(+)**	Regulation committee	0.00665
Sector committee(-)	**Regulation committee(+)**	Africa	0.00712
Oceania(-)	**Asia(+)**	None	0.00924

Table 9: Any combination of two distributions that rejected the null-hypothesis of the Mann-Whitney test on an alpha of 0.01.

As can be seen in Table 9, the insurance industry is significantly more negative compared to both the regulation industry and the consulting industry. However, to strengthen or weaken this conclusion on these industrial differences, the distribution of the continent of operation is shown for each of these three industries in Figure 12.

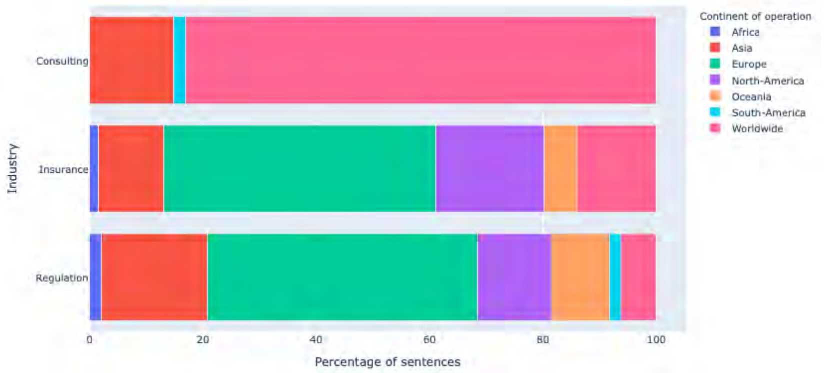


Figure 12: The distribution (%) of the continents of operation per three different industries: consulting, insurance, and regulation.

As can be seen in Figure 12, consulting has a very different distribution in continents of operation compared to insurance. More than 80% of the consulting organizations are operating worldwide, while less than 20% of the insurance organizations are operating worldwide. Also, almost 50% of the insurance organization are operating in Europe only, while there are no consulting organization that are operating in Europe only. This means that most of the consulting distribution is determined by the distribution of the organizations operating worldwide, while a big part of the insurance distribution is determined by the distribution of the organizations operating in Europe. It is highly possible that the significant difference in sentiment between companies operating worldwide and companies operating in Europe, as can be seen in Table 9, indirectly takes part into the conclusion of the difference between insurance and consultancy.

On the other hand, the insurance industry almost has a similar distribution of continents of operation compared to the one of the regulation industry. The percentages of each of the continents of operation are very similar for both the distributions, making the significant difference in sentiment between the insurance distribution and the regulation distribution a reliable difference.

6.2 Thorough analysis

The previous section answers the question "Can we find any statistically different opinions on IFRS 17 from the organizations concerned?" with a definite yes since we found 9 statistically different opinions between geographical and industrial groups concerned. However, it is not yet clear what the organizations are talking about when they are either positive or negative. This section will focus on a more thorough analysis to find the topics of sentiment within the significantly different opinions and thereby tries to answer the follow-up questions from Section 2.3: "Can we find the topics that organizations are talking about when they are either very negative or positive?", and "Can we find some of the most negative and positive comments on IFRS 17?".

Performing thorough analysis on all 9 statistically different opinions will take a tremendous number of pages. Therefore, only one analysis will be done in this Section as an example. The example analysis will be done on the two distributions of insurance and regulation, which are significantly different according to the Mann-Whitney test. This combination is chosen due to the fact that the data imbalance analysis, done at the end of the previous section, shows no imbalance between those two distribution, which strengthens the conclusion of a presence of significant difference. Also, this is an interesting analysis for EY specifically since EY itself is has to deal with both regulators and insurers (most of its clients are insurers).

The distributions of the sentiment scores of the insurance industry and the regulation industry are plotted in the Figure 13. The median of the insurance industry is below -0.5 and the median of the regulation industry is above 0.0. Since it is statistically proven that the samples are not from the same distribution, it is possible to say that the regulation industry is more positive than the insurance industry.

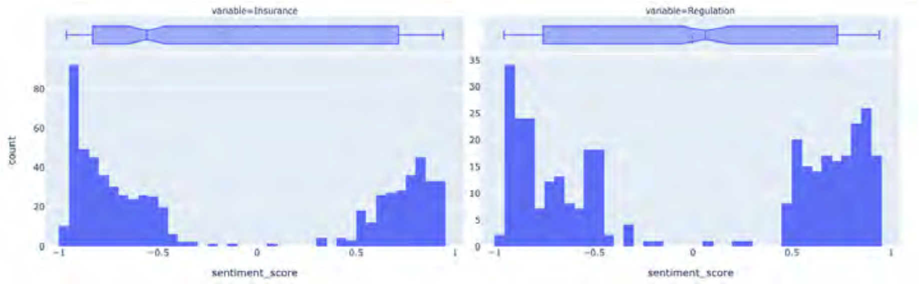


Figure 13: The distribution of the sentiment scores of the insurance industry (left) and the regulation industry (right).

6.2.1 Word clouds of sentiment

To answer the research question ”What are the topics that organizations are talking about when they are significantly more negative/positive ?”, we will make use of word clouds to gain insights. Word clouds show the most common words within a specific text. To be able to gain insights into one distribution of text, two word clouds are created for this group. One word cloud contains the most common words in the negatively labeled sentences and the other word cloud contains the most common words in the positively labeled sentences. This way, it is possible to see what a group is (mostly) talking about when it is either negative or positive.

Before being able to create a word cloud, a pre-process has to be performed on the input text. This pre-process is necessary to separate each word, remove meaningless words like ’and’ and ’or’, and recognize words like ’walk’ and ’walking’ as the same words. The methods used for this are respectively called tokenization [36], removing stop words, and lemmatization [37]. The algorithm to pre-process a text containing one or more sentences before being able to create a word cloud is coded as shown in Algorithm 1.

Algorithm 1 Pre-processing text for a word cloud

```
Input: text
Output: list of tokens $t$
$t \leftarrow$ empty list
for sentence in text do
    for token in Tokenize(sentence) do
        token $\leftarrow$ lowercase(token)
        if token is not a number then
            if token is not a stop word then
                if length(token) $>1$ then
                    token $\leftarrow$ Lemmatize(token)
                    $t$.append(token)
return $t$
```

Algorithm 1 returns a list of tokens that is the correct input for single word clouds in Python documentation. The word cloud generator in Python takes in a list of all the subsequent words, then counts the number of times that each word occurs in the list and shows the most frequent words in the word cloud, where the more frequent words have a larger size. In this case, the word clouds come in pairs: one negative word cloud and one positive word cloud, where it does not add valuable insights if both word clouds show similar words. Unconditional frequent words, which are frequent in both negative and positive labeled sentences, are therefore removed before used as an input for the word cloud generator in Python documentation.

The four word clouds originating from the positive and negative sentences written by the insurance industry and the regulation industry are shown in Figure 14. The two word cloud on the left represent the most common words in separately the negatively (top) and positively (bottom) labeled sentences from the comment letters written by the insurance industry. The two word cloud on the left represent the most common words in separately the negatively (top) and positively (bottom) labeled sentences from the comment letters written by the regulation industry.

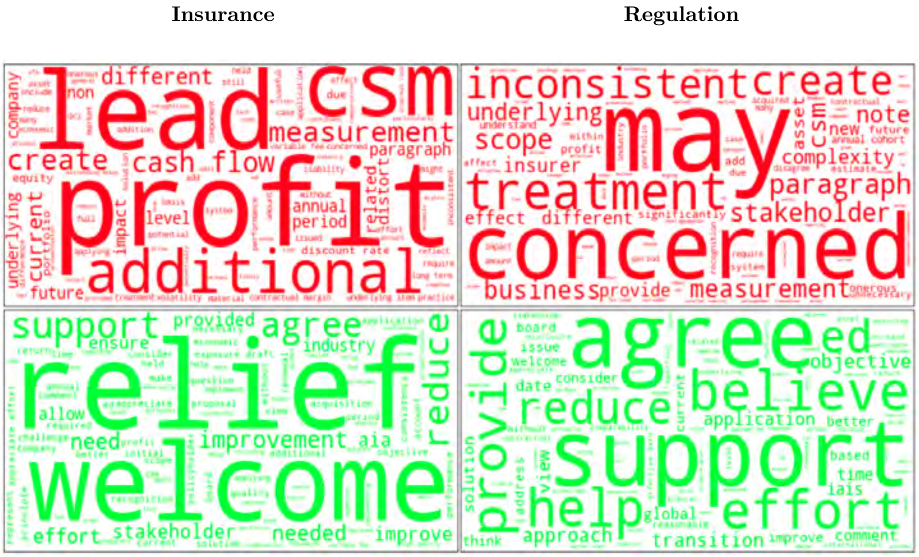


Figure 14: Word clouds of the most frequent words in the negatively (top) and positively (bottom) labeled sentences originating from the comment letters written by insurers (left) and regulators (right).

The word clouds from Figure 14 show a difference in vocabulary between the insurers and the regulators in their sentimental sentences. It is important to look at the context in which this vocabulary is used, as it for example differs if the word 'may' is referring to the expression of possibility or to the month May. The context became clear after quickly checking the sentences containing these vocabulary, which are not presented in this thesis due to the large quantity. The regulators are quite general in their positive opinions as they 'agree' and 'support' the amendments. Their negative expressions show they are 'concerned' that future negative situations 'may' rise, and worried that some parts are 'inconsistent'. The fact that 'inconsistent' is a frequent word is interesting as regulators generally check presence of a level playing field and stability in the sector. A chance of inconsistency would then be worrying, especially since inconsistency is present right now (IFRS 4) and IFRS 17 is issued to gain more comparability. The insurers are more subject specific as they have an issue with 'CSM', which is a key component calculating the expected insurance contract 'profit' (Section 2). The insurers 'welcome' certain amendments like the delay of the effective date, and other amendments that offer 'relief', which indicates they experience challenges with implementing the standard. The subjects that insurers and regulators are sentimental on can perhaps give value to EY if further looked into by specialists, for example on what aspects to the CSM are of specific concern, for what specific challenge the relief is welcomed, and what parts of the amendments may be inconsistent according to regulators and why.

6.2.2 Most sentimental paragraphs

The word clouds give a very broad idea of the opinions of the groups, but not anything specific. A different method can be used to present specific examples of negative or positive pieces of text. This method is performed by calculating the most sentimental paragraphs of all the selected comment letters. The input data is the text without any cleaning, so the neutral sentences included. To be specific, the moving average of the sentiment scores is calculated for any subsequent 9 sentences. The index that corresponds to the highest or lowest average sentiment score is the last index of the 9 most sentimental sentences. The number of 9 sentences can however easily be changed but this number is chosen since it is approximately the length of a paragraph. As a direct example, this paragraph is also 9 sentences long. The most negative and positive 9 sentences of the comment letters of the regulation industry and the insurance industry are shown separately below, where the negatively labeled sentences are shown in red and the positively labeled sentences are shown in green.

Most **negative** 9 sentences of the **regulation** industry:

Writer = Accounting Standards Board (AcSB) [Canada]; Document = LINDAMEZONAccountingStandardsBoardAcSB_0; Pages = 19

We observe that it is common for subsidiaries around the world not to report on a stand-alone interim basis. Therefore, we think that the IASB should clarify paragraph B137 of IFRS 17 so that a subsidiary can apply the same accounting estimates used in the preparation of the parent's consolidated interim financial statements in accordance with IAS 34 to the preparation of its annual financial statements.

Accounting treatment of contracts acquired during the settlement period in a business combination

Our stakeholders have also expressed concerns with the IASB's rationale noted in paragraphs BC206 to BC207 of the Basis for Conclusions on the Exposure Draft that supports the removal of the exception in paragraph 17 of IFRS 3 Business Combinations. The removal of the exception results in entities having to account for insurance contracts acquired during their settlement period as new insurance contracts. This means that the liability for incurred claims acquired will have to be reclassified as a liability for remaining coverage, even though the insured event covered by the insurance contract has already occurred. The effect of this accounting treatment for insurance contracts acquired during the settlement period will inappropriately gross up the amount of insurance service revenue and insurance service expenses recognized. As a result, the entity's financial performance will be distorted. As well, this accounting treatment will add significant complexity and increase costs to the financial reporting processes.

Most **positive** 9 sentences of the **regulation** industry:

Writer = Accounting Standards Board (AcSB) [Canada]; Document = LINDAMEZONAccountingStandardsBoardAcSB_0; Pages = 2

Our views

Given the stage of global implementation, we support the criteria developed by the IASB to assess which possible amendments could be justified at this time, without causing undue delay in implementation efforts. We strongly support completion of this process, leading to a final IFRS 17, because the standard is expected to benefit the global capital markets. Financial reporting under this new insurance contract standard will help users better understand current and future results of operations and financial position of entities that issue insurance contracts and will serve to improve global comparability. We commend the IASB's efforts to respond to the implementation concerns that stakeholders raised. We think that the proposed amendments in this Exposure Draft are a positive step to help entities effectively implement IFRS 17 by addressing key challenges that entities have encountered thus far and help manage implementation costs.

Effective date proposal

In our view, a common global adoption date of IFRS 17 is critical to the success of transition by supporting the move to greater global comparability. It is in the global capital markets' best interests that entities adopt this new standard at the same time to ease the transition challenges for users, preparers, auditors, and regulators around the world. We stand ready to join the IASB and other national standard setters in getting this standard finalized and achieving a common global adoption date.

Most **negative** 9 sentences of the **insurance** industry:

Writer = Korea Life Insurance Association (KLIA); Document = SusieLeeKoreaLifeInsuranceAssociationKLIA_0; Pages = 2, 3

○ As of the end of June 2019, the fixed-guaranteed insurance book accounts for 41% of the total premium reserves which takes up the most of liability reserves. Its proportion has declined by only 3 percentage points for the last 4.5 years from 44%.

○ In terms of policies offering the fixed-guarantee of no less than 5%, its book with remaining duration for 20 years or more account for 74% of the total premium reserves. This structure is extremely vulnerable to mark-to-market liabilities evaluation.

○ The burden imposed by a large amount of premium reserves for high fixed-guarantee insurance policies is expected to continue for long, which in turn will serve as an impediment to the introduction of IFRS 17 under which insurers have to measure the liabilities by market value.

The world is suffering from the low interest rates environment. Unlike European countries, Korea in particular has experienced a sharp drop* in interest rates while further reduction is projected given the circumstances home and abroad.
* IRs of Government bonds in Korea: 2.469% → 1.948% → 1.596% → 1.172% The base interest rate will drop further in the second half of 2019.

○ In amidst of the low interest rates landscape, the introduction of IFRS 17 which requires insurers to report insurance liabilities to market value, not book value, will lead to rapid increase in liabilities and decrease in capital.

Most **positive** 9 sentences of the **insurance** industry:

Writer = Samsung Life Insurance; Document = HAEINLEESamsungLifeInsurance_0; Pages = 1

As a country of full IFRS adoption, South Korea recognizes and appreciates the work of the IASB and the significance of the IFRSs.

We would like to appreciate this opportunity to comment on and express support for the Exposure Draft ED/2019/4 Amendments to IFRS 17.

The proposed amendments increase practicability and preserve the principles of IFRS 17. The amendments capture the IASB's efforts to reflect the industry's opinions, as can be shown in the significant consideration of the discussions at the TRG.

Overall, Samsung Life supports the proposed amendments. In particular, many support the newly introduced concept of investment-return service, which permits them to recognize the CSM based on the service patterns of economic substance.

However, we would like to note a few amendments that might be worthy of further consideration. They can be interpreted in different ways depending on the reader. We believe that our opinions may help improve the Standard, which are described in Appendix that follows.

The pieces of text above give an insight into the opinions of singular organizations on specific subjects. Both the most negative and positive pieces of text of the regulation industry are written by the Accounting Standards Board of Canada (AcSB), suggesting that they are relatively sentimental. Their most negative part shows that their stakeholders do not agree with the removal of a specific exception since it will have negative effects on the financial performance of entities. The most positive part of the AcSB is less specific as it shows their overall support on IFRS 17, however asking kindly for a response on the concerns of the stakeholders. In the most negative piece of text of the insurance industry, the Korea Life Insurance Association is giving reasons why more time is required for the preparation of IFRS 17. It is stating facts on the vulnerability of the insurance liability structure and the sharp drop in interest rates in Korea. The most positive piece of text of the insurance industry however does not state any specific subject. It shows Samsung Life Insurance being kind and writing a positive introduction of their comment letter, stating that they recognize and appreciate the work of the IASB, appreciate the opportunity to comment, support the amendments, however believe that their opinions may help improve IFRS 17. These opinions following in their Appendix probably contain more subject specific information. Finding out that the second most negative 9 sentences are written by Samsung Life Insurance, indicates that these opinions are more negative. These 9 sentences are shown below, where Samsung Life Insurance expects more than 7.5 million euro's of sunk costs to arise due to the amendments, and states unclarity of the effective date of IFRS 17 at the time of writing the comment letter. Both the examples of the insurance and the regulation industry show a certain specificity on the negative subjects and a generality on the positive subjects. This perhaps indicates that negatively labeled text contains more information compared to positively labeled text.

Second most **negative** 9 sentences of the **insurance** industry:

Writer = Samsung Life Insurance; Document = HAEINLEESamsungLifeInsurance_0; Pages = 11

However, we had to change our systems because of the proposed amendments, and this has already resulted in a significant delay in our schedule. Also, with the deferral of the effective date from 2021 to 2022, we are expecting more than 7.5 million EUR of sunk costs to arise, related to system maintenance. If the IASB chooses to amend IFRS 17 further, this can disrupt the implementation processes underway, and can disadvantage most who have sought to implement IFRS 17 on a timely basis.

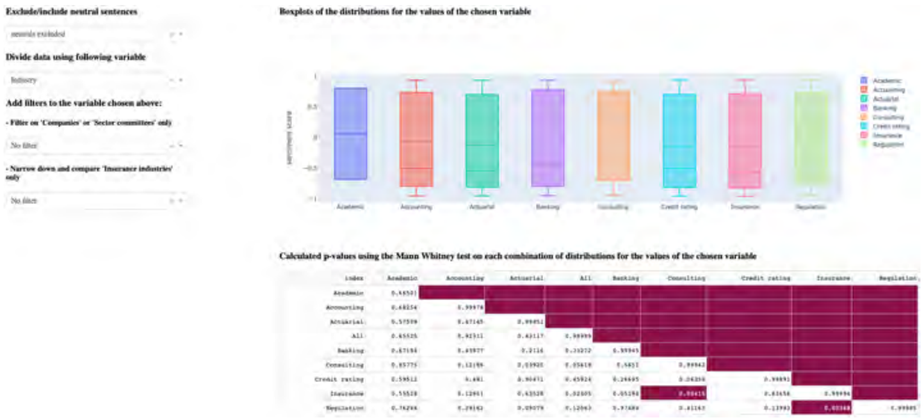
Effective date of IFRS 17 is important for IFRS 9 as well. Banks and securities firms are concerned and dissatisfied that only insurers are exempt from IFRS 9. They have questioned the fairness of automatically extending temporary exemption from IFRS 9 because of the deferral of IFRS 17.

IFRS 17 is the first comprehensive and international accounting standard for insurance contracts and it is important that the effective date is aligned globally. If the implementation date of IFRS 17 differs by jurisdictions, it would disrupt comparability of financial information among different jurisdictions. It is only likely that there would be complaints and concerns from those that implement IFRS 17 first.

6.3 Dashboard

The analysis in the previous section (Section 6.2) is only performed on the significant difference between the insurance industry and the regulation industry. This is only 1 of the 9 significant differences in opinion statistically recognized by using the Mann-Whitney test on an alpha of 0.01 as shown in Table 9. When deciding to take a higher alpha of, for instance, 0.05, there are 24 significant differences in opinion statistically recognized as can be seen in Table 29. Doing such a number of analyses at once is impractical as it will quickly result in an increasing number of print results of the code and pages in the report. A more practical alternative is to code the Python documentation that makes it possible to manually change the variables of the two distributions that are desired to be compared. However, this can only be done by someone that is in possession of the code and has enough knowledge of data science to work with Python to change the variables in the code. IFRS 17 specialists usually do not have this specific programming knowledge and the data science specialists on the other hand do not have the specialized knowledge on IFRS 17 to do thorough analysis. Since both these parties are on the other side of the spectrum, this creates a gap that cannot be filled without any tool. MIT Sloan and BCG: "Among the 90%

of the companies that have made at least some investment in AI, fewer than 2 out of 5 report obtaining any business gains from AI in the past three years" [38]. To unlock the value of AI, data science has to be a team sport. It would be of great value to build an environment where IFRS 17 specialist without any knowledge of Python programming are able to do the analyses themselves. For that reason, a dashboard has been build using Python Dash. The start position of the dashboard looks like shown in Figure 15.



Exclude/include neutral sentences

neutrals excluded

Divide data using following variable

Industry

Add filters to the variable chosen above:

- Filter on 'Companies' or 'Sector committees' only

No filter

- Narrow down and compare 'Insurance industries' only

No filter

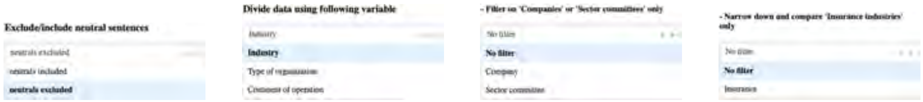
Boxplots of the distributions for the values of the chosen variable

Calculated p-values using the Mann Whitney test on each combination of distributions for the values of the chosen variable

index	Academic	Accounting	Actuarial	All	Banking	Consulting	Credit rating	Insurance	Regulation
Academic	0.46501								
Accounting	0.68254	0.99978							
Actuarial	0.57509	0.67145	0.99951						
All	0.65525	0.92511	0.42117	0.99999					
Banking	0.67184	0.45977	0.2116	0.33272	0.99945				
Consulting	0.85775	0.12189	0.03925	0.05618	0.5851	0.99942			
Credit rating	0.59512	0.481	0.90471	0.45924	0.26695	0.26356	0.99891		
Insurance	0.55528	0.12951	0.42528	0.02505	0.05194	0.00415	0.83658	0.99996	
Regulation	0.76266	0.29142	0.09079	0.12063	0.97684	0.41163	0.13983	0.03368	0.99985

Figure 15: Start position of the dashboard.

The left side of the dashboard shows multiple drop-downs that allow the visuals on the right side of the dashboard to change. The first drop-down makes it possible to include or exclude the neutral sentences from the data as can be seen in Figure 16a. The default value is the option to exclude the neutrals since the neutral sentences disrupted the data as explained in Section 5.2. The second drop-down makes it possible to choose the variable that the data will be divided on. As can be seen in Figure 16b, the default value is the industry variable, which can easily be changed to either the type of organization or the continent of operation, depending on the desired analysis of the user. The drop-down(s) after the second one are dependent on the specific variable chosen in the second drop-down. In this case, the third drop-down makes it possible to filter the industry variable on companies or sector committees only, as can be seen in Figure 16c. The fourth drop-down makes it possible to filter, separately or additionally, the industry variable on organizations from insurance industries only, as can be seen in Figure 16d.



Exclude/include neutral sentences

neutrals excluded

neutrals included

neutrals excluded

(a)

Divide data using following variable

Industry

Industry

Type of organization

Continent of operation

(b)

- Filter on 'Companies' or 'Sector committees' only

No filter

No filter

Company

Sector committee

(c)

- Narrow down and compare 'Insurance industries' only

No filter

No filter

Insurance

(d)

Figure 16: The drop-down menu's of the start position of the dashboard.

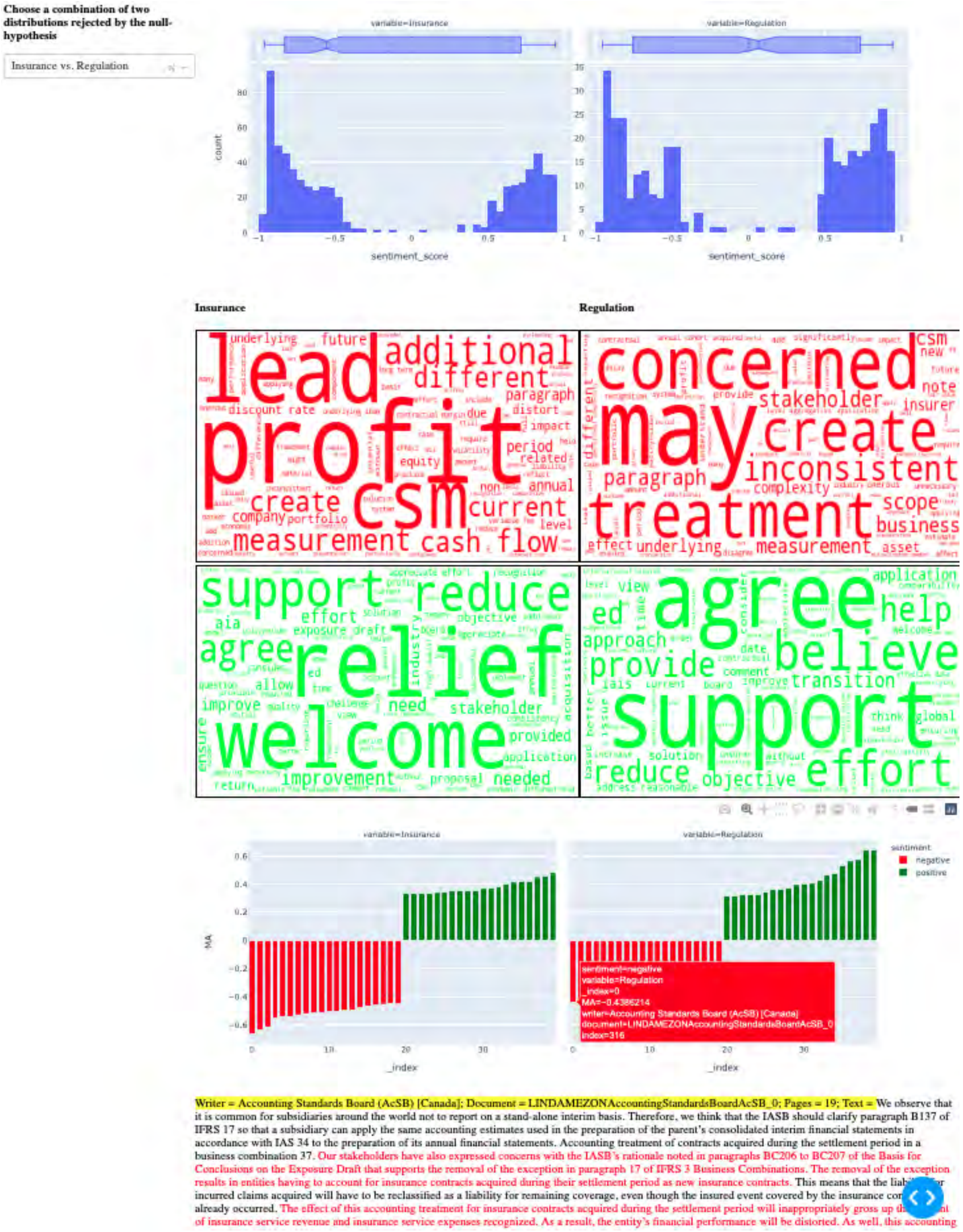
When taking a look at the start position of the dashboard as shown in Figure 15, the visuals on the right side show the box plots of the distributions for the different labels of the chosen variable and the table with the calculated p-values using the Mann-Whitney test on every possible combination of labels of the chosen variable. The p-values that are lower than 0.01 are colored in purple. This way, the table shows if any significant differences exist (tested on an alpha of 0.01) and the box plots show which one is relatively positive and which one is relatively negative. In the case of choosing the industry variable to be compared, the two combinations of labels that reject the null hypothesis are consulting versus insurance and insurance versus regulation. When scrolling down on the dashboard, it shows an additional drop-down menu that makes it possible to choose any combination of two distributions of the chosen variable that rejects the null-hypothesis with a p-value below 0.01. The drop-down menu always looks like shown in Figure 17a. The options of this drop-down menu are similar to the the combinations that got a purple color in the table. In this case, the options of the drop-down menu are therefore consulting versus insurance and insurance versus regulation as can be seen in Figure 17b.

Choose a combination of two distributions rejected by the null-hypothesis
Select a combination
Choose a combination of two distributions rejected by the null-hypothesis
Select a combination
Consulting vs. Insurance
Insurance vs. Regulation

(a) (b)

Figure 17: The drop-down menu that shows the combinations of variable labels of the current table visually available on the dashboard that got rejected by the null hypothesis with a p-value below 0.01.

When choosing for the combination of insurance versus regulation, the dashboard visualizes the analysis as shown in Figure 18, which is similar to the analysis done in Section 6.2. However, instead of only the most positive and negative 9 sentences of the text, it is now possible to check the top 20 most positive and the top 20 most negative 9 sentences. The dashboard shows the distribution, the positive and negative word clouds, and a bar graph representing the top 20 most negative and the top 20 most positive 9 sentences of text, for both groups separately. This bar graph is interactive as it is possible to click on any bar and the corresponding piece of text will appear below. In this case, the most negative piece of text of the regulation industry is selected and shown on the bottom. This is the piece of text written by the AcSB as was also shown in Section 6.2.2.



Choose a combination of two distributions rejected by the null-hypothesis

Insurance vs. Regulation

sentiment=negative
variable=Regulation
_index=0
MA=−0.4386214
writer=Accounting Standards Board (AcSB) [Canada]
document=LINDAMEZONAccountingStandardsBoardAcSB_0
index=316

Writer = Accounting Standards Board (AcSB) [Canada]; Document = LINDAMEZONAccountingStandardsBoardAcSB_0; Pages = 19; Text = We observe that it is common for subsidiaries around the world not to report on a stand-alone interim basis. Therefore, we think that the IASB should clarify paragraph B137 of IFRS 17 so that a subsidiary can apply the same accounting estimates used in the preparation of the parent's consolidated interim financial statements in accordance with IAS 34 to the preparation of its annual financial statements. Accounting treatment of contracts acquired during the settlement period in a business combination 37. Our stakeholders have also expressed concerns with the IASB's rationale noted in paragraphs BC206 to BC207 of the Basis for Conclusions on the Exposure Draft that supports the removal of the exception in paragraph 17 of IFRS 3 Business Combinations. The removal of the exception results in entities having to account for insurance contracts acquired during their settlement period as new insurance contracts. This means that the liab r incurred claims acquired will have to be reclassified as a liability for remaining coverage, even though the insured event covered by the insurance con already occurred. The effect of this accounting treatment for insurance contracts acquired during the settlement period will inappropriately gross up th t of insurance service revenue and insurance service expenses recognized. As a result, the entity's financial performance will be distorted. As well, this accounting

Figure 18: Second part of the dashboard showing a thorough analysis on the selected combination of two distributions.

7 Topic modeling

Throughout the process of this research, a lot of methods have been researched and tested that did not present the sensible results useful for this project. Before finding the data set of comment letters specifically used for this research, attempts have been made to create a model that calculates the sentiment over time. Combining this with topic modeling would result in a model that calculates the sentiment over time per topic, similar to the research done by De Winter and Van Dijk [3]. The data set of De Winter and Van Dijk consists of a million news articles containing only a few sentences. The data online available on IFRS 17 is structurally however very different since there are only a limited number of articles but multiple pages long. It soon became clear that this available data on IFRS 17 was not suitable for a model that calculates sentiment over time per topic. Therefore, the research focused on creating a topic model applicable on singular large articles. Topic modeling is a machine learning technique that statistically analyses text data to determine clusters that represent the abstract "topics" for a set of documents. A known topic model is Latent Dirichlet Allocation (LDA) [7], also used by De Winter and Van Dijk to calculate the sentiment per topic [3]. Applying this model on a large document however did not show any sensible results, since it created one cluster with approximately 90% of the words included. When testing the model on an online data set existing of thousands of financial news articles on stocks, it performed successfully as it showed clear clusters on for example medicine, commodities, politics, etc. This shows again that the structure of the data set does not fit a model, since the LDA model cannot be performed on a single large document but only on lots of small articles.

The dedication to find different topics in the documents resulted in a different attempt based on the article "Analysis of Markov Influence Graphs" written by Berkhout and Heidergott [8]. The article analyses node graphs by calculating a distance between each node. The distance is the average number of steps that needs to be taken before a node returns to itself or to another node. This distance of a node with itself is called the mean recurrence time and the distance of a node with another node is called the mean first passage time [39]. The distances between each node can be used to cluster the nodes. The paper of Berkhout and Heidergott applies this method on some example data sets like social networks, where the nodes represent the people in the network. In the case of this research, the nodes represent the words present in a document. The distances between each of the occurring words will then be clustered, meaning that the words closer to each other will be clustered together. Two steps need to be taken before being able use the calculations that result in a matrix containing the mean recurrence and the mean first passage times as presented in the paper by Berkhout and Heidergott. The first step is pre-processing the raw text data. The pre-process is coded as shown in Algorithm 2.

Algorithm 2 Pre-processing raw text data

```
    Input: text
    Output: list of tokens $t$
$t \leftarrow$ empty list
for sentence in text do
    negation $\leftarrow$ False            ▷ A negation word is a word like no, not, non
    $t$.append('<s>')                   ▷ '<s>' symbolizes the start of a sentence
    for token in Tokenize(sentence) do
        token $\leftarrow$ lowercase(token)
        if token is not a number then
            if token is a negation word then
                negation $\leftarrow$ True
                go to next iteration
            if token is not a stopword then
                if length(token) >1 then
                    token $\leftarrow$ lemmatize(token)
                    if negation is True then
                        $t$.append('!'+token)
                    else
                        $t$.append(token)
    $t$.append('</s>')                  ▷ '</s>' symbolizes the end of a sentence
return $t$
```

The input of Algorithm 2 is raw text data and the output is a list containing the occurring words in order. The words are changed to lower case, stop words are removed, numbers are removed, and every word is inflected to its root form using lemmatization. Also during this process, tokens are added to the list representing the start ('<s>') and the end ('</s>') of a sentence, and negation words are changed to '!' and added to the beginning of the next token. A negation word is a word like 'no', 'not' or 'non', which are words that occur in the stop words dictionary. Since all stop words are removed, the algorithm would otherwise change two subsequent words like 'not good' to 'good'. However, the meaning of 'good' is the exact opposite of the meaning of 'not good'. Therefore 'not' is remembered and '!' is added to the word 'good' resulting in '!good'. An example of an input and a corresponding output of Algorithm 2 is shown in Example 1.

Example 1.
Input: text = ”EFRAG is not aware of any European insurer having taken a firm commitment to early apply the Standard. Finally, EFRAG notes that IFRS 17 requires a presentation of restated comparative information when applying the Standard for the first time.”
Output: t = [’<s>’, ’efrag’, ’!aware’, ’european’, ’insurer’, ’taken’, ’firm’, ’commitment’, ’early’, ’apply’, ’standard’, ’</s>’, ’<s>’, ’finally’, ’efrag’, ’note’, ’ifrs’, ’requires’, ’presentation’, ’restated’, ’comparative’, ’information’, ’applying’, ’standard’, ’first’, ’time’, ’</s>’]

After the pre-process, the data is suitable to be converted into a transition probability matrix, or Markov chain as coded in Algorithm 3. This matrix shows the conditional probability $P(word_{column}|word_{row})$ that one word follows another for every combination of words. Since the sum of the probabilities of all possible events has to be equal to 100%, the sum of each row of a Markov chain is always equal to one. After generation of the markov chain, the algorithm deletes the rows and the columns that are indexed with the tokens that symbolize the start and end of a sentence (’<s>’ and ’</s>’). These tokens prevented the algorithm from also counting the first word of a sentence to follow the last word of the previous sentence, but they are useless for the steps coming after this. Subsequently to this step, the algorithm normalizes each row to make the sum of each row equal to one again.

Algorithm 3 Transition probability matrix

```
    **Input**: list of tokens $t$
    **Output**: matrix $P$
$u \leftarrow$ list containing unique string values from $t$
$P \leftarrow$ matrix of zeros with index and columns set to $u$
**for** $i$ in 0,1,2,...,(length($t$)−1) **do**
    idx $\leftarrow t[i]$
    col $\leftarrow t[i+1]$
    $P[\text{idx}, \text{col}] \leftarrow P[\text{idx}, \text{col}] + 1$
$P \leftarrow$ remove indices and columns with string values ’<s>’ and ’</s>’
$P \leftarrow$ normalize each row of $P$
**return** $P$
```

The input of Algorithm 3 is the output list of tokens from Algorithm 2 and the output of Algorithm 3 is a matrix representing the Markov chain on each word occurring in the text. An example of an input and a corresponding output of Algorithm 3 is shown in Example 2.

Example 2.
Input: t = ['<s>', 'efrag', 'aware', 'european', 'insurer', 'taken', 'firm', 'commitment', 'early', 'apply', 'standard', '</s>', '<s>', 'finally', 'efrag', 'note', 'ifrs', 'requires', 'presentation', 'restated', 'comparative', 'information', 'applying', 'standard', 'first', 'time', '</s>']

Output: $$P = \begin{array}{r|ccccc|} & \text{efrag} & \text{!aware} & \text{finally} & \text{note} & \text{...} \\ \hline \text{efrag} & 0 & 0.5 & 0 & 0.5 & \text{...} \\ \text{!aware} & 0 & 0 & 0 & 0 & \text{...} \\ \text{finally} & 1 & 0 & 0 & 0 & \text{...} \\ \text{note} & 0 & 0 & 0 & 0 & \text{...} \\ \text{...} & \text{...} & \text{...} & \text{...} & \text{...} & \text{...} \end{array}$$

Since the data is converted into a Markov chain P, it is now possible to calculate the matrix M containing the mean first passages times with the calculations as shown in the paper from Berkhout and Heidergott [8]. The calculations are as follows:

$$\Pi_P = \lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} P^n \tag{1}$$

$$D_P = \left(I - P + \Pi_P\right)^{-1} - \Pi_P \tag{2}$$

$$M = \Big(I - D_P + \overline{11}^{\top} \cdot \mathrm{dg}\left(D_P\right)\Big) \cdot \mathrm{dg}\left(\Pi_P\right)^{-1} \tag{3}$$

where dg(A) results from A by setting off-diagonal entries to zero.

The calculation can only be done when P is irreducible [40]. If P is not irreducible, a small number is added to every entry in the matrix. This results in the coding of the mean first passage times as shown in Algorithm 2.

Algorithm 4 Mean first passage times

```
Input: matrix $P$
Output: matrix $M$
if $P$ is not irreducible then
    $P \leftarrow (1-\theta)P + \theta I$    ▷ $I \leftarrow$ matrix containing ones; $\theta \leftarrow$ small number
$\Pi(P) \leftarrow$ equation (1)    ▷ $\Pi \leftarrow$ matrix with stationary distributions of $P$
$D(P, \Pi) \leftarrow$ equation (2)    ▷ $D \leftarrow$ deviation matrix of $P$
$M(\Pi, D) \leftarrow$ equation (3)    ▷ $M \leftarrow$ mean first passage times of $P$
return $M$
```

The input of the calculation is the Markov chain and the output is the matrix containing the mean first passage times. An example of an input and a corresponding output of the calculations shown above is shown in Example 3. The input P of this example is the output of Example 2.

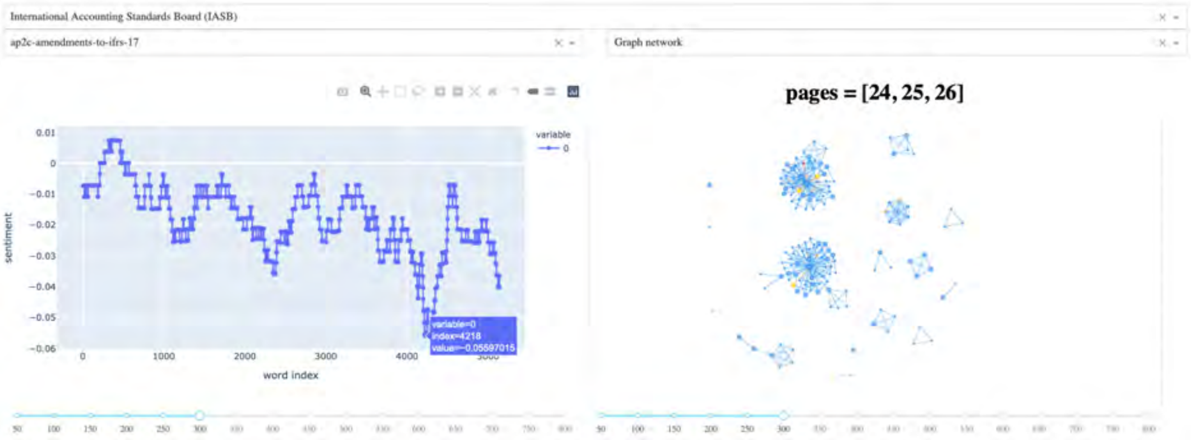
Example 3.

Input: $$P = \begin{array}{r|ccccc|} & \text{efrag} & \text{!aware} & \text{finally} & \text{note} & ... \\ \text{efrag} & 0 & 0.5 & 0 & 0.5 & ... \\ \text{!aware} & 0 & 0 & 0 & 0 & ... \\ \text{finally} & 1 & 0 & 0 & 0 & ... \\ \text{note} & 0 & 0 & 0 & 0 & ... \\ ... & ... & ... & ... & ... & ... \end{array}$$

Output: $$M = \begin{array}{r|ccccc|} & \text{efrag} & \text{!aware} & \text{finally} & \text{note} & ... \\ \text{efrag} & 75.5 & 38.75 & 150.0 & 38.75 & ... \\ \text{!aware} & 74.5 & 75.50 & 149.0 & 75.50 & ... \\ \text{finally} & 1.0 & 39.75 & 151.0 & 39.75 & ... \\ \text{note} & 74.5 & 75.50 & 149.0 & 75.50 & ... \\ ... & ... & ... & ... & ... & ... \end{array}$$

The example output shows high numbers like 151 meaning that it takes on average 151 steps to reach the next word. The list of tokens used to create this matrix, which is the output of Example 1, only contains 24 words. When the mean first passage time is larger than the number of words in the text, this means that these words are not linked. The only reason that a distance exists is due to the added matrix containing small numbers at each entry since P was irreducible.

The words can now be clustered by considering the mean first passage times and the mean recurrence times as the distance between words. Attempts of finding sensible clusters have been time consuming, taking weeks to months. A multitude of clustering methods have been applied, however they all did not show sensible results. These were clustering methods such as k-medoids [9], k-means [10], DBSCAN [11], Louvain [12], agglomerative hierarchical clustering [13], and Greedy Modularity Maximization [14]. Perhaps there were too many dimensions to effectively cluster, which is why all cluster methods were also performed in combination with methods for dimensionality reduction like PCA [15] and t-SNE [16], making it also possible to visualize the data set before and after clustering. This however did not contribute to the quality of the cluster results. The clustering methods often showed one large cluster and only sometimes one or more side-clusters containing only a handful of words. Either the clustering methods needed more parameter tuning or the data set is not suited for these methods. However, the clustering methods took a long time to run on a large document. Therefore, a custom method is used, which creates clusters by only keeping the n edges with the lowest values of mean first passage time and then plotting the graph network that remains. Figure 19 shows an application of this method in an interactive environment generated with Python Dash again. The graph network is created using the *Pyvis* Python package [41].



International Accounting Standards Board (IASB)

ap2c-amendments-to-ifrs-17

Graph network

pages = [24, 25, 26]

variable=0
index=4218
value=−0.05597015

Figure 19: Dashboard showing an interactive analysis of a single document, using a sentiment plot and a graph network.

The dashboard in Figure 19 shows an analysis of one of the three documents written by the IASB as a summary of all the comment letters they received. As the input data is converted to a list of words with Algorithm 2, the more successful FinBERT model from Section 4.2 can not be used in this case since this model can only classify sentences. Therefore, the plot on the left shows the moving average of the sentiment given to the words using the word list model as explained in Section 4.1. The moving average window can be adjusted below the plot using the slider and, as can be seen, is currently set on 300 words. When clicking on a data point in the sentiment plot, the dashboard creates a graph network on the right. Figure 19 shows the graph network created using the 300 subsequent words corresponding to the lowest average sentimental value at index 4218. The pages corresponding to these 300 words are shown above the graph network, which are in this case pages 24, 25, and 26. The value in the slider below the graph network represent the number of edges in the graph network. The higher the value on the slider, the more connections and consequently less clusters. Reversely, the lower the value on the slider, the less connections and consequently more clusters. Figure 20 shows a zoom-in on the three largest clusters shown in the graph network in Figure 19.

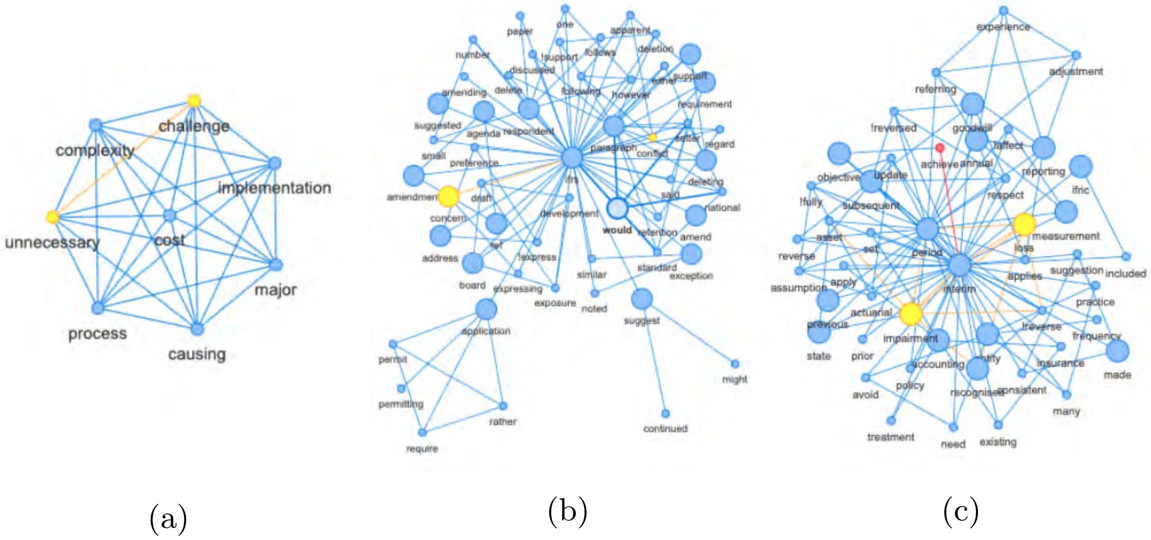


Figure 20: The three largest clusters from the graph network shown in Figure 19.

Some words in the three clusters from Figure 20 show a different color. The blue colored nodes represent neutral words, the yellow colored nodes represent negative words, and the red colored nodes represent positive words. A positive word having a red color does not make sense intuitively, but the colors were given automatically and cannot be changed manually. Words that appear in the text only once are programmed to be represented by a smaller node. This since these words will always have a small distance towards at least one word. The large nodes appear at least two times in the text. The clusters shown in Figure 20a, 20b, and 20c represent subjects about respectively cost, IFRS, and interim period. All three subjects have a small distance towards some negatively labeled words. For further analyses, it would be interesting to have a look on pages 24, 25, and 26 and see why the writer is specifically negative on these subjects.

Such a dashboard makes it possible to do topic and sentiment analyses on one document at a time. This dashboard is therefore not applicable for the specific research of finding statistically significant differences between groups. However, the tool is applicable for analysing singular large documents. For example, if an academic or an employee has to read a document with a lot of pages, say 200, this person can then save time by using this dashboard to know where to find the most sentiment and what it is about. Therefore, the research on the methodology of topic modeling did not result in a successfully useful method for this specific research, but it did result in a by-product that can be useful for further or other research.

8 Concluding remarks and further research directions

This thesis defined research questions to give direction to the goal, where the main research question is "Can we find statistically significant opinions on IFRS 17 from the organizations concerned?", and the follow-up questions are "Can we find the topics that organizations are talking about when they are either very negative or positive?" and "Can we find some of the most negative and positive comments on IFRS 17?". As a conclusion, the main research question can be answered affirmatively as this thesis shows groups of organizations that statistically have a significant different opinion compared to each other. Since the main research question can be answered affirmatively, the follow-up questions have also been researched using an interactive environment applicable for thorough analysis. To be precise, the results are based on a data set that originated from an event where the IASB issued amendments on IFRS 17 in *ED/2019/4 Amendments to IFRS 17* and asked the public to comment. This was the chance for any organization concerned to express their opinion towards the IASB, resulting in comment letters coming from 123 different organization from all over the world and from any industry concerned. After an extremely time consuming process of searching for a data set, this data set was chosen since the quantity of the comment letters and the probability of this data set containing sentiment made it suitable for further research. To give the data set an extra dimension and to make it possible to compare different groups of organizations, each organization was manually labeled by head office, continent of operation, type of organization, industry type, and insurance type. To be able to quantify and analyse the data, the PDF files had to be extracted to Python. Six different PDF to text converters were tested in Python that all showed a few unsuccessful text extractions. The most successful text extractor turned out to be *fitz* that extracted 98 out of the 123 different organization. The quantification of the data is done by calculating the sentiment per sentence. Two data science models were validated using the validation set of Financial PhraseBank from Malo et al. [5], where *FinBERT for sentiment classification* [4] scored significantly higher compared to the model based on the *Sentiment Lexicon* by Loughran and McDonald [2], with respectively an accuracy of 0.97 versus an accuracy of 0.65 (on the validation set with an agreement level of 100%). Therefore, *FinBERT for sentiment classification* was used to calculate the sentiment score per sentence of each comment letter. This created the opportunity to compare the sentiment distributions of different geographical and industrial groups. Since the distributions are not even close to being normally distributed, the non-parametric Mann-Whitney test with the null-hypothesis "the two samples come from the same distribution" is used to compare two groups. The organizations copy large parts of the standards and questions asked by the IASB before giving an answer, which results in an excessive number of neutral sentences. To ensure comparison of sentiment instead of neutrality, the neutral labeled sentences were removed as part of data cleaning. The p-values that were calculated using the Mann-

Whitney test for every relevant combination of two groups are shown in the Appendix (Section 10.2-10.5). As a result, there are 9 different pairs of groups that have a p-value lower than 0.01 and therefore a significantly different opinion relative to each other, as shown in Table 9. For example, both the consultants and the regulators separately seem to be significantly more positive compared to the insurance industry. This means that the main research question is answered affirmatively, and therefore further analysis has been done to give answer to the follow-up questions. Performing thorough analysis on all 9 statistically different opinions would however take a tremendous number of pages, which is why only one analysis is shown in this thesis. The example analysis is done on the two distributions of insurance and regulation, because the analysis of data imbalance showed no imbalance between those two distribution meaning there is no possibility of this significant difference being an indirect consequence of another. To give an answer to the first follow-up question, word clouds were created that give insights in the most frequent words per sentiment label of both the insurers and the regulators, as shown in Section 6.2.1. Insurers are frequently negative on 'CSM', which is a key component of IFRS 17, and frequently positive on the delay of the effective date offering 'relief', indicating they experience challenges with implementing the standard. The regulators were frequently negative using the words 'concerned' and 'inconsistent', perhaps indicating that they are concerned that (part of) the standards are inconsistent. This would be worrying since inconsistency is present right now (IFRS 4) and IFRS 17 is issued to gain more comparability. For a more specific idea of the context and to answer the second follow-up question, the most sentimental paragraphs (9 sentences) were calculated and presented, as shown in Section 6.2.2. The specific examples of context show a certain specificity on the negative examples and a generality on the positive examples, perhaps indicating that negatively labeled text contains more information compared to positively labeled text. To easily perform analysis on all other statistically different opinions as well, this research provides an easy-to-use dashboard, created with Python Dash. This tool makes it possible for IFRS 17 specialists without any knowledge of Python programming to do the analyses themselves, unlocking the value of AI and therefore making data science a team sport. This research therefore shows a methodology that provides results both applicable in literature and businesses such as EY.

Besides the previous discussed methodology and tool, this thesis also delivers a by-product. Before finding the specific data set used for this research, a lot of methods and models have been tested on different data sets. Due to the limited but long articles online available on IFRS 17, a significant part of this research has been dedicated to finding a successful topic model applicable on singular documents. After testing a dozen of different methods, the only method showing sensible results is a custom clustering method on the mean recurrence times and mean first passage times (from the paper of Berkhout and Heidergott [8]) of the words in the document. This topic model is implemented into a (second) easy-to-use dashboard that analyses singular documents by interactively showing the most sentimental parts and their corresponding topic clusters. This (second) tool can be useful for purposes other than the specific goal of this research.

Further research directions can therefore focus on deploying the by-product on large documents. This can be done both academic favored and business favored since both academics and business employees have to deal with large documents of (financial) text once in a while. Besides just deploying the by-product, further research can also focus on testing and perhaps validating this tool. The clusters shown in the topic model seemed to make sense, especially compared to the other topic modeling methods. However, the complexity of the subject in combination with the lack of knowledge on IFRS 17 kept it a challenge to be absolutely sure on the logic of the (combinations of) terms in the clusters. The topic model can be tested by either letting an IFRS 17 specialist look at it or by using a different document that is about an easy and understandable subject. Using a document as input while having knowledge on the subject will give more insights into the potential function or dysfunction of the product. The tool allows to select, for example, the most negative part in the sentiment plot. The topic clusters corresponding to this part will then be shown together with the page numbers. The topic clusters can then be validated by checking these pages in the document, reading if and why the writer is specifically negative, and see if this matches with the topic clusters shown on the dashboard. Other further research on the by-product can focus on ways to improve the tool, for example, to include an analysis on the centrality scores of each word. A centrality score is a measure of the influence, and therefore the importance, of a node in a network [42]. Calculating the centrality scores will make it possible to show a hierarchy in words, where the most connected words are on top, either for the whole document or per cluster. Centrality can either be calculated using a simple centrality score such as Betweenness [43] or a more complex centrality score such as PageRank [44], which is also used by successful companies like Google. Since the centrality score is not automatically correlated with the frequency of a word, this can perhaps give new insights and shift attention.

As discussed above, the by-product shows a lot of potential for further research. The main research part of this thesis however shows a lot of potential for further research directions as well. This further research can for instance focus on improving the methodology and tool that is used to answer the research questions. This tool currently shows statistical testing, distributions, word clouds, and the most sentimental paragraphs. An idea to improve the tool is to add an analysis by finding the 'most common' negative or positive sentence of a group. As an example, let's take all the negative sentences of the insurance industry. The 'most common' sentence can then be found by calculating the similarity between every negative sentence and every other negative sentence in the insurance industry. Each sentence will then have a number of similarity scores equal to the number of total negative sentences in the insurance industry minus one. Taking the average of these similarity scores for each sentence will result in a score that represents the magnitude of a sentence being similar to all the other sentences. The sentence with the highest score has the most in common with all the other sentences. This sentence will therefore be a representative of the 'most popular' negative opinion of the insurance industry. Practice will however tell if the results of this method will be sensible and gain actual knowledge on the 'most popular' opinion. Similarity between two sentences can be calculated using a variety of string matching algorithms [45]. The NLP model BERT can also be deployed to find text similarity by using these texts as an input for the pre-trained BERT without fine-tuning and then comparing the output embeddings. Li et al. [46] however argue that the BERT embeddings poorly capture semantic meaning of sentences, and therefore propose a method that exploits performance of semantic similarity. Their open-source code can thus be used to find quality similarity scores between sentences. Besides this potential improvement and addition to the tool of the main research part, other further research directions can focus on expansion of the thorough analysis of the results. This thesis only performs a thorough analysis on the opinions of the insurers and the regulators, which has been done by a Business Analytics Masters student with a lack of knowledge on IFRS 17. It will be more relevant to let a IFRS 17 specialist do these analyses for a few hours, not just on the opinions of insurers versus regulators, but on all 9 statistically significant differences in opinion. Maybe even all 24 statistically significant differences in opinion, which rejected the null-hypothesis of the Mann-Whitney test with an alpha of 0.05, can be analysed. This creates an opportunity to validate the functioning and user-friendliness of the tool as the specialist is able to give feedback and simultaneously gathers valuable knowledge on IFRS 17. Validation will make the results of the tool more valuable. The tool is now specifically deployed for comment letters on IFRS 17. However, further research can focus on this tool being deployed for a different relevant subject such as the pension agreement in the Netherlands. This is currently a hot topic and relevant for the actuarial department of EY as pension funds are part of their client base. When comment letters on the pension agreement are available, the methodology of this research can be used to provide a similar tool that can statistically compare the opinions of the organizations concerned. This can help EY gain knowledge on the way their clients are thinking, meaning more valuable and goal oriented consult.

9 Bibliography

References

[1] T. Loughran and B. McDonald, "When is a liability not a liability? textual analysis, dictionaries, and 10-ks," *The Journal of Finance*, vol. 66, no. 1, pp. 35–65, 2011.

[2] T. Loughran and B. Mcdonald, "Measuring firm complexity," *SSRN Electronic Journal*, 2020.

[3] J. de Winter and D. van Dijk, "Sentimentsindicator op basis van financieel economisch nieuws (sentimentindicator based on financial economic news)," *Economisch Statistische Berichten*, forthcoming (2021).

[4] D. Araci, "Finbert: Financial sentiment analysis with pre-trained language models," *CoRR*, vol. abs/1908.10063, 2019.

[5] P. Malo, A. Sinha, P. Takala, P. Korhonen, and J. Wallenius, "Financialphrasebank-v1.0," 07 2013.

[6] N. Nachar, "The mann-whitney u: A test for assessing whether two independent samples come from the same distribution," *Tutorials in Quantitative Methods for Psychology*, vol. 4, 03 2008.

[7] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," in *Advances in Neural Information Processing Systems* (T. Dietterich, S. Becker, and Z. Ghahramani, eds.), vol. 14, MIT Press, 2002.

[8] J. Berkhout and B. F. Heidergott, "Analysis of markov influence graphs," *Operations Research*, vol. 67, no. 3, pp. 892–904, 2019.

[9] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for k-medoids clustering," *Expert Systems with Applications*, vol. 36, no. 2, Part 2, pp. 3336–3341, 2009.

[10] J. A. Hartigan and M. A. Wong, "A k-means clustering algorithm," *JSTOR: Applied Statistics*, vol. 28, no. 1, pp. 100–108, 1979.

[11] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, p. 226–231, AAAI Press, 1996.

[12] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, p. P10008, Oct 2008.

[13] D. Müllner, "Modern hierarchical, agglomerative clustering algorithms," *CoRR*, vol. abs/1109.2378, 2011.

[14] A. Clauset, M. E. J. Newman, and C. Moore, “Finding community structure in very large networks,” *Physical Review E*, vol. 70, Dec 2004.

[15] A. Maćkiewicz and W. Ratajczak, “Principal components analysis (pca),” *Computers Geosciences*, vol. 19, no. 3, pp. 303–342, 1993.

[16] L. van der Maaten and G. Hinton, “Viualizing data using t-sne,” *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 11 2008.

[17] International Accounting Standards Board, “Why global accounting standards?.” https://www.ifrs.org/use-around-the-world/why-global-accounting-standards/. Accessed: 2021-23-10.

[18] International Accounting Standards Board, “About us.” https://www.ifrs.org/about-us/. Accessed: 2021-23-10.

[19] A. Aarzen and T. Mourik, “Ifrs 4 insurance contracts,” *Maandblad Voor Accountancy en Bedrijfseconomie*, vol. 79, pp. 10–18, 01 2005.

[20] International Accounting Standards Board, “Ifrs 17 insurance contracts.” https://www.ifrs.org/issued-standards/list-of-standards/ifrs-17-insurance-contracts.html/content/dam/ifrs/publications/html-standards/english/2021/issued/ifrs17/. Accessed: 2021-23-10.

[21] A. Fleischmann and J. Hirz, “The ifrs17 guide for the perplexed actuary,” *Mathematics eJournal*, 2018.

[22] W. Yousuf, J. Stansfield, K. Malde, N. Mirin, R. Walton, B. Thorpe, J. Thorpe, C. Iftode, L. Tan, R. Dyble, and et al., “The ifrs 17 contractual service margin: a life insurance perspective,” *British Actuarial Journal*, vol. 26, p. e2, 2021.

[23] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, “A convolutional neural network for modelling sentences,” *CoRR*, vol. abs/1404.2188, 2014.

[24] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, “Recurrent neural network based language model,” vol. 2, pp. 1045–1048, 01 2010.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *CoRR*, vol. abs/1706.03762, 2017.

[26] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” *CoRR*, vol. abs/1810.04805, 2018.

[27] National Institute of Standards and Technology (U.S.), “Reuters Corpora,” 2018.

[28] J. Howard and S. Ruder, “Fine-tuned language models for text classification,” *CoRR*, vol. abs/1801.06146, 2018.

[29] M. Kraus and S. Feuerriegel, “Decision support from financial disclosures with deep neural networks and transfer learning,” *CoRR*, vol. abs/1710.03954, 2017.

[30] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, (New Orleans, Louisiana), Association for Computational Linguistics, June 2018.

[31] H. Keselman, A. R. Othman, R. R. Wilcox, and K. Fradette, “The new and improved two-sample t test,” *Psychological Science*, vol. 15, no. 1, pp. 47–51, 2004. PMID: 14717831.

[32] K. P. Balanda and H. L. Macgillivray, “Kurtosis: A critical review,” *The American Statistician*, vol. 42, no. 2, pp. 111–119, 1988.

[33] F. J. M. Jr., “The kolmogorov-smirnov test for goodness of fit,” *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[34] S. S. Shapiro and M. B. Wil, “An analysis of variance test for normality (complete samples),” *Biometrika*, vol. 52, pp. 591–611, 12 1965.

[35] R. D’Agostino and E. S. Pearson, “Tests for departure from normality. Empirical results for the distributions of b2 and b1,” *Biometrika*, vol. 60, pp. 613–622, 12 1973.

[36] J. Webster and C. Kit, “Tokenization as the initial phase in nlp,” pp. 1106–1110, 01 1992.

[37] V. Balakrishnan and L.-Y. Ethel, “Stemming and lemmatization: A comparison of retrieval performances,” *Lecture Notes on Software Engineering*, vol. 2, pp. 262–267, 01 2014.

[38] S. Ransbotham, S. Khodabandeh, R. Fehling, and B. L. andD. Kiron, “Winning with ai,” *MIT Sloan Management Review and Boston Consulting Group*, 2019.

[39] Z. Zhang, A. Julaiti, B. Hou, H. Zhang, and G. Chen, “Mean first-passage time for random walks on undirected networks,” *The European Physical Journal B*, vol. 84, pp. 691–697, 2011.

[40] T. J. Sheskin, “Computing mean first passage times for a markov chain,” *International Journal of Mathematical Education in Science and Technology*, vol. 26, no. 5, pp. 729–735, 1995.

[41] G. Perrone, J. Unpingco, and H. Lu, "Network visualizations with pyvis and visjs," *CoRR*, vol. abs/2006.04951, 2020.

[42] G. Lawyer, "Understanding the influence of all nodes in a network," *Scientific reports*, vol. 5, p. 8665, 03 2015.

[43] L. C. Freeman, "A Set of Measures of Centrality Based on Betweenness," *Sociometry*, vol. 40, pp. 35–41, Mar. 1977.

[44] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," pp. 161–172, 1998.

[45] N. Singla and D. Garg, "String matching algorithms and their applicability in various applications," *International Journal of Soft Computing and Engineering (IJSCE)*, vol. 1, 01 2012.

[46] B. Li, H. Zhou, J. He, M. Wang, Y. Yang, and L. Li, "On the sentence embeddings from pre-trained language models," *CoRR*, vol. abs/2011.05864, 2020.

10 Appendix

10.1 Complete set of labels

op_continent	**Nr.**	**head_office**	**Nr.**	**org_type**	**Nr.**	**industry**	**Nr.**	**insur_branch**	**Nr.**
Africa	7	Australia	7	Academic	2	Academic	2	Composite	22
Asia	27	Austria	2	Company	38	Accounting	17	Life	12
Europe	37	Belgium	6	Regulation committee	24	Actuarial	13	Non-life	5
North-America	9	Brazil	1	Sector committee	59	Banking	11	Reinsurance	3
Oceania	7	Canada	9			Consulting	10		
South-America	3	Chile	1			Credit rating	2		
Worldwide	33	China	8			Insurance	42		
		Colombia	1			Regulation	24		
		Dubai	1						
		France	8						
		Germany	11						
		India	2						
		Indonesia	1						
		Ireland	1						
		Italy	1						
		Japan	4						
		Kenya	1						
		Malaysia	2						
		Netherlands	1						
		New Zealand	1						
		Nigeria	1						
		Poland	1						
		Singapore	3						
		Slovenia	1						
		South Africa	3						
		South Korea	5						
		Spain	3						
		Sweden	1						
		Switzerland	3						
		Tanzania	1						
		Thailand	1						
		United Kingdom	21						
		United States	6						
		Zimbabwe	1						

Table 10: The complete set of labels given to each comment letter and the corresponding number of documents that got this label.

10.2 Geographical comparisons

	Africa	Asia	Europe	North-America	Oceania	South-America
Asia	0.44572					
Europe	0.75588	0.26525				
North-America	0.8673	0.076	0.25694			
Oceania	0.34978	0.00924	0.04113	0.33537		
South-America	0.3221	0.49013	0.32979	0.23778	0.0693	
Worldwide	0.20491	0.30524	0.01313	0.00604	0.00065	0.75381

Table 11: The p-values of the Mann-Whitney test performed on every combination of continent of operation.

Filtered on Organization type = Company

	Africa	Asia	Europe	North-America	South-America
Asia	0.44041				
Europe	0.99412	0.0364			
North-America	0.80786	0.14091	0.47692		
South-America	0.93265	0.69784	0.91025	0.97023	
Worldwide	0.32021	0.54528	0.00216	0.01343	0.63304

Table 12: The p-values of the Mann-Whitney test performed on every combination of continent of operation (filtered on companies only).

Filtered on Organization type = Sector committee

	Africa	Asia	Europe	North-America	Oceania
Asia	0.1681				
Europe	0.28866	0.39111			
North-America	0.57034	0.35561	0.60701		
Oceania	0.89189	0.03845	0.08488	0.39598	
Worldwide	0.12678	0.84481	0.26038	0.27079	0.02926

Table 13: The p-values of the Mann-Whitney test performed on every combination of continent of operation (filtered on sector committees only).

Filtered on Organization type = Regulation committee

	Africa	Asia	Europe	North-America	Oceania	South-America
Asia	0.01113					
Europe	0.0199	0.56103				
North-America	0.02469	0.95048	0.67242			
Oceania	0.00665	0.20625	0.08795	0.35252		
South-America	0.20134	0.09358	0.23217	0.17607	0.04305	
Worldwide	0.2515	0.06499	0.14848	0.09992	0.0139	0.97967

Table 14: The p-values of the Mann-Whitney test performed on every combination of continent of operation (filtered on regulators only).

10.3 Industrial comparisons

	Academic	Accounting	Actuarial	Banking	Consulting	Credit rating	Insurance
Accounting	0.68254						
Actuarial	0.57509	0.47145					
Banking	0.67194	0.43977	0.2116				
Consulting	0.85775	0.12186	0.03925	0.0.5651			
Credit rating	0.59512	0.481	0.90471	0.26695	0.06356		
Insurance	0.55528	0.12961	0.63528	0.05198	0.00415	0.83656	
Regulation	0.76244	0.29162	0.09079	0.97484	0.41165	0.13993	0.00348

Table 15: The p-values of the Mann-Whitney test performed on every combination of industry.

Filtered on Organization type = Company

	Banking	Consulting	Credit rating
Consulting	0.8149		
Credit rating	0.62013	0.06356	
Insurance	0.50078	0.0167	0.96186

Table 16: The p-values of the Mann-Whitney test performed on every combination of industry (filtered on companies only).

Filtered on Organization type = Sector committee

	Accounting	Actuarial	Banking
Actuarial	0.47145		
Banking	0.46631	0.23807	
Insurance	0.12862	0.58269	0.07139

Table 17: The p-values of the Mann-Whitney test performed on every combination of industry (filtered on sector committees only).

10.4 Organizational comparisons

	Academic	Company	Regulation committee
Company	0.6642		
Regulation committee	0.76244	0.12957	
Sector committee	0.61625	0.86785	0.06366

Table 18: The p-values of the Mann-Whitney test performed on every combination of organization type.

Filtered on Continent of operation = Africa

	Company	Regulation committee
Regulation committee	0.04544	
Sector committee	0.833	0.00712

Table 19: The p-values of the Mann-Whitney test performed on every combination of organization type (filtered on organizations operating in Africa only).

Filtered on Continent of operation = Asia

	Academic	Company	Regulation committee
Company	0.22731		
Regulation committee	0.24465	0.82692	
Sector committee	0.32893	0.92333	0.80702

Table 20: The p-values of the Mann-Whitney test performed on every combination of organization type (filtered on organizations operating in Asia only).

Filtered on Continent of operation = Oceania

	Regulation committee
Sector committee	0.91764

Table 21: The p-values of the Mann-Whitney test performed on every combination of organization type (filtered on organizations operating in Oceania only).

Filtered on Continent of operation = Europe

	Company	Regulation committee
Regulation committee	0.0058	
Sector committee	0.03765	0.17124

Table 22: The p-values of the Mann-Whitney test performed on every combination of organization type (filtered on organizations operating in Europe only).

Filtered on Continent of operation = North-America

	Company	Regulation committee
Regulation committee	0.28241	
Sector committee	0.50588	0.63099

Table 23: The p-values of the Mann-Whitney test performed on every combination of organization type (filtered on organizations operating in North-America only).

Filtered on Continent of operation = South-America

	Company
Regulation committee	0.36193

Table 24: The p-values of the Mann-Whitney test performed on every combination of organization type (filtered on organizations operating in South-America only).

Filtered on Continent of operation = Worldwide

	Company	Regulation committee
Regulation committee	0.1554	
Sector committee	0.71402	0.14904

Table 25: The p-values of the Mann-Whitney test performed on every combination of organization type (filtered on organizations operating Worldwide only).

10.5 Insurance type comparisons

	Composite	Life	Non-life
Life	0.20953		
Non-life	0.44138	0.94218	
Reinsurance	0.74852	0.91209	0.87461

Table 26: The p-values of the Mann-Whitney test performed on every combination of insurance industries.

Filtered on Organization type = Company

	Composite	Life
Life	0.06829	
Non-life	0.48217	0.77273

Table 27: The p-values of the Mann-Whitney test performed on every combination of insurance industries (filtered on companies only).

Filtered on Organization type = Sector committee

	Composite	Life	Non-life
Life	0.7157		
Non-life	0.30735	0.52665	
Reinsurance	0.76088	0.98707	0.94363

Table 28: The p-values of the Mann-Whitney test performed on every combination of insurance industries (filtered on sector committees only).

10.6 Rejected null-hypothesis

Mann-Whitney test input					
The two distributions		**Filtered on**	**p-value**	**< 0.01**	**< 0.05**
Oceania(-)	**Worldwide(+)**	None	0.00065	✔	✔
Europe(-)	**Worldwide(+)**	Company	0.00216	✔	✔
Insurance(-)	**Regulation(+)**	None	0.00348	✔	✔
Consulting(+)	**Insurance(-)**	None	0.00415	✔	✔
Company(-)	**Regulation committee(+)**	Europe	0.00580	✔	✔
North-America(-)	**Worldwide(+)**	None	0.00604	✔	✔
Africa(+)	**Oceania(-)**	Regulation committee	0.00665	✔	✔
Regulation committee(+)	**Sector committee(-)**	Africa	0.00712	✔	✔
Asia(+)	**Oceania(-)**	None	0.00924	✔	✔
Africa	**Asia**	Regulation committee	0.01113	✘	✔
Europe	**Worldwide**	None	0.01313	✘	✔
North-America	**Worldwide**	Company	0.01343	✘	✔
Oceania	**Worldwide**	Regulation committee	0.01390	✘	✔
Consulting	**Insurance**	Company	0.01670	✘	✔
Africa	**Europe**	Regulation committee	0.01990	✘	✔
Africa	**North-America**	Regulation committee	0.02469	✘	✔
Oceania	**Worldwide**	Sector committee	0.02926	✘	✔
Asia	**Europe**	Company	0.03640	✘	✔
Company	**Sector committee**	Europe	0.03765	✘	✔
Asia	**Oceania**	Sector committee	0.03845	✘	✔
Actuarial	**Consulting**	None	0.03925	✘	✔
Oceania	**Europe**	None	0.04113	✘	✔
Oceania	**South-America**	Regulation committee	0.04305	✘	✔
Company	**Regulation committee**	Africa	0.04544	✘	✔

Table 29: Any combination of two distributions that rejected the null-hypothesis of the Mann-Whitney test on separately an alpha of 0.01 and 0.05.