When Firms Open Up: Identifying Value Relevant Textual Disclosure Using simBERT

September 5, 2022

Abstract

By introducing *simBERT*, a novel semantically sensitive similarity measure for textual data, we find that international annual reports contain value relevant information that is not timely priced by investors. We measure the value relevance of international corporate disclosures by constructing a portfolio that is long in stocks with a low- and short in stocks with a high level of semantically new information. Such a portfolio yields a highly significant yearly abnormal return of 8.52%. We observe a higher value relevance of textual disclosure in developed countries, which we trace back to stricter securities laws standards. Our findings thus indicate that tighter regulation promotes the disclosure of value relevant accounting information. We further find evidence that analysts update their earnings forecasts and recommendations in accordance with textual changes in firm reports. This suggests that analysts contribute to market efficiency by conveying qualitative information from accounting statements to the public.

Keywords: Textual Analysis, Value Relevance, Securities Laws, Annual Reports, Disclosure, BERT

1. Introduction

Assessing the value relevance of corporate disclosures remains a central aspect within the accounting literature. A common method to assess the value relevance of firm disclosure is to relate stock returns to quantitative data like cash flow, income or balance sheet figures (see Barth et al., 2008). Textual data however remains relatively unexplored, especially in an international setting. A major reason is that automatically extracting new information from (multilingual) text is not straightforward. Textual data has to be transformed into a numerical representation to process it in an automated manner. Depending on the efficiency of the transformation process, information might be lost during this process and thus will not be accessible to researchers and investors alike. However, if researchers disregard textual data, the obtained view on the value relevance of corporate disclosure might be incomplete.

It seems likely that these obstacles to extract value-relevant information from text are amplified for international markets due to language barriers, different disclosure requirements, and a lacking standardized format for firm reports. While there exists a predetermined annual report structure in the US, the so-called 10-K form, which is submitted in a HTML file to the Securities and Exchange Commission (SEC), firms used to disclose information via PDF file internationally.[1] To study international corporate disclosures, researchers have to rely on the non-harmonized PDF files which arguably complicates the automatic extraction of specific firm report information. Moreover, as firms change the format of these PDFs over time, firms sometimes rephrase sentences in this process. By doing so, they invoke document changes that do not convey semantically new information, reducing the accuracy of simple but widely used word-based similarity measures like bag-of-words *(bow)*, which cannot filter out these changes.

Against this background, we investigate whether international firm reports contain value relevant information. We collect and process more than 300,000 annual reports from over 18,000 non-US firms across 30 countries from Bloomberg. By suggesting and applying a novel context-aware similarity measure, we find that investors overlook value relevant information within international annual reports. We examine a strategy that is long in *non-changers*, i.e. firms that disclose only little semantically new information, and short in *changers*, i.e. firms

[1] Recently, there has been some progress in the harmonization of corporate disclosure outside the US, for instance the European Single Electronic Format (ESEF) which requires European firms to disclose in a XHTML format since 2020. For more information, see: https://www.esma.europa.eu/sections/european-single-electronic-format. However, accessing these files remains challenging because the European Electronic Access Point (EEAP), a central database similar to Edgar in the US, is not yet available.

that disclose a large amount of semantically new information. Controlling for various risk factors, this strategy delivers a highly significant abnormal return of up to 71 monthly basis points, indicating that text in annual reports indeed contains value relevant information.

Our paper is related to the well-noticed cross-country analysis of Lang and Stice-Lawrence (2015), who show that IFRS adoption is associated with an increase in disclosure quality, as measured by the increase in firm report length and a decrease in boilerplate[2]. Within this paper, we extend their analysis by testing if international firm reports contain value relevant information. We further investigate whether stricter securities laws standards as discussed in La Porta et al. (2006) are associated with a higher value relevance. Due to our multilingual framework, we are able to reduce the self-selection bias mentioned by Lang and Stice-Lawrence (2015, p.4) that occurs if the analysis is restricted to English firm reports.[3]

To test whether annual reports contain value relevant information, we calculate the abnormal return of a portfolio that is long in *non-changers* and short in *changers*. If we find a significant difference in the performance of *changers* in comparison to *non-changers* that is not explainable by any commonly known pricing factor, we may conclude that corporate disclosures contain value relevant textual information.[4] Our approach is inspired by Cohen et al. (2020) who investigate whether investors fully price new firm report information in the US. However, instead of using a bag-of-words approach to proxy the level of new information within a document, we propose *simBERT*, a novel document similarity measure that leverages a pre-trained sentence-transformer model[5] to estimate the share of semantically new information within a document. Using this measure, we are able to filter out irrelevant document changes.

Our measure might be allocated to the area of Semantic Textual Similarity (STS) which comprises models that measure the "semantic equivalence between two blocks of text" (Chandrasekaran and Mago, 2021, p.1). It rests on a sentence-transformer model, which is a modified version of BERT (Bidirectional Transformers for Language Understanding) that returns semantically sensitive vector representations for a given input text. These representations may be compared via cosine similarity, a measure that is intuitively similar to a correlation coefficient.

[2]Boilerplate is described as non-informative, non-firm-specific information that is included to avoid legal disputes (Lang and Stice-Lawrence, 2015).

[3]The reason is that firms from non-English speaking countries decide themselves whether they provide an English report or not.

[4]For more information on the portfolio construction, see section 3.2.

[5]In particular, we obtain sentence representations by applying a multilingual sentence-transformer model called *paraphrase-multilingual-mpnet-base-v2 (Reimers and Gurevych, 2020)*.

BERT itself is capable of learning bidirectional representations from unlabeled text. Using so-called *attention mechanisms*, the model learns to identify important parts of a text and how they connect to other parts (Devlin et al., 2018).

In the field of Natural Language Processing *(NLP)*, this particular area of research has experienced substantial contributions within the last years. With the increase in computational power, large corporations like Google, Facebook and others constantly publish larger and yet more powerful language models (see e.g. Devlin et al., 2018; Liu et al., 2019; Brown et al., 2020; Chowdhery et al., 2022). Given that these language models are able to quantify the semantic meaning of sentences, they serve as an ideal foundation for similarity measures that can control for meaningless document changes which is a major advantage over established word-based similarity measures like bag-of-words.

To illustrate the performance of the multilingual sentence embeddings that we use to identify new information, let us consider various modifications of a shortened sentence taken from the letter of the chairman of Anglo American' 2020 annual report, *"Lockdowns in certain countries put additional pressure on our mining operations."* On the one hand, the cosine similarity between this sentence and a semantically similar one like *"Shutdowns in specific countries were a challenge for our mining operations."* is 0.83 and thus draws the correct conclusion that the sentences are semantically similar. A bag-of-words *(bow)* approach yields a quite low cosine similarity of 0.5 even though these two sentences are highly similar from a semantic perspective.

On the other hand, the cosine similarity that we obtain when comparing the vector representation for our original sentence and the one for a sentence like *"Lockdowns in certain countries did not put additional pressure on our mining operations."* is lower (0.70) which correctly indicates that these sentences have a different meaning. Instead, the bow approach yields a cosine similarity of 0.93 which implies a high similarity. The example shows that the vectors obtained from the multilingual model can be superior to word frequency vectors as they are capable to capture the semantic meaning of a sentence.

In contrast to the *bow* approach, where documents are represented at the word level, *simBERT* calculates the similarity of two documents at the sentence level. For each sentence in a given document, we identify the most similar sentence in the previous document of the same firm by calculating pairwise cosine similarities of the sentence embeddings obtained from the pre-trained language model. We then average the maximum cosine similarities, i.e. the cosine similarities of the pairs of most similar sentences identified before, to obtain *simBERT*.

Our similarity measure *simBERT* is superior in comparison to word-based methods in multiple dimensions. First, it may control for irrelevant document changes. If a firm decides to restructure and rephrase its firm report, it might induce document changes that do not contain any semantically new information. Since a *bag-of-words* approach can not differentiate between semantically new information and irrelevant document changes, rephrased sentences would likely lead to an imprecise measure of document similarity. Second, while a *bag-of-words* approach may only compare the similarity of documents that are written in the same language, *simBERT* may be applied to documents that are written in any of the fifty languages supported by the underlying language model. Third, in contrast to *bag-of-words*, which is upwards biased for document pairs with a large set of words by construction, *simBERT* yields unbiased similarity estimates even for the largest documents. Finally, we show that our measure substantially outperforms a *bag-of-words* approach on a labeled dataset of US business descriptions. By subdividing US business sections, we find that *simBERT* is able to correctly detect the missing part of a firm's business description with an accuracy of 82.00%, whereas *bag-of-words* achieves no more than 65.67% accuracy.

Our analyses rest on three hypotheses. At the report-level, we first hypothesize that controlling for document changes that do not convey semantically new information should be more important in the context of non-US firm reports that lack a fixed report structure. The reason is that a change in report structure often coincides with rephrased sentences that do not convey semantically new information. Word-based approaches that do not consider the context of words may not control for these changes and thus might be biased. As a result, we expect to see a larger difference in the amount of value relevant information, as estimated by *bag-of-words* and *simBERT*, outside the US.

Our second hypothesis argues that cross-country differences in the value relevance of annual reports should be related to the country's institutional framework, in particular its regulatory environment and securities laws standards. For instance, La Porta et al. (2006) find that disclosure requirements, liability standards for investor protection, and other regulatory aspects influence the development of financial markets. We hypothesize that the same factors will also determine the performance of our text-based long-short strategy, as companies should have much more incentives to transparently communicate value-relevant news in their annual statements if the (legal) consequences of failing to do so are more severe. This hypothesis is supported by Lang and Stice-Lawrence (2015) who find that "textual attributes are predictably

associated with regulation and incentives for more transparent disclosures"(p.1). Thus, we expect that there should be more value relevant information disclosed in reports of firms that operate in countries with a stricter regulatory environment.

Our third hypothesis explores whether financial analysts influence to what extent value relevant information is timely priced. Bradshaw (2011) argues that our understanding of the role of analysts for capital markets is still limited, despite an extensive literature on their behavior and performance. More recently, Guo et al. (2020) find that analyst recommendations are biased towards overpriced stocks, while Azevedo and Müller (2021) assert that this effect is much less pronounced for many international markets. We expect analysts to be attentive readers of financial statements, who should be able to understand and act on value-relevant information that is disclosed in text changes of annual reports. This implies that analysts should primarily downgrade their earnings forecasts and recommendations for *changers*. Doing so, they may promote market efficiency by making capital market participants aware of bad news early. Therefore, we expect a stronger and faster price response, and hence a weaker return for stocks with extensive analyst coverage compared to stocks with little or no analyst coverage.[6]

To investigate the above-mentioned hypotheses, we use our international dataset of annual reports. These reports are usually available in the official language of a country, however for some countries, our dataset contains English reports instead.[7] In total, our dataset contains more than 300,000 international annual reports from over 18,000 firms across 30 countries from 1995 until 2021. For the US, we collect 185,000 annual reports (10-K files) for more than 8,000 firms from the SEC. We regress the return differences between a portfolio that is long (short) in *non-changers* (*changers*) on six commonly known factors to control for various differences in the composition of the two portfolios. We then treat the constant (alpha) of the regression as unexpected return and test its significance using a t-test. If we find a significantly positive alpha, we may conclude that annual reports indeed contain value relevant information.

We start the empirical analysis by simulating US calendar time portfolios based on *bow* and *simBERT*. We go long in *non-changers* and short in *changers* and obtain results that are

[6]It is obvious that we cannot test if analysts actually read the annual reports. It is possible that they gather the same information which is communicated in annual reports from other events or documents like conference calls, earlier quarterly reports, press releases, or private meetings with company representatives. This, however, does not invalidate the hypothesis that analysts could be contributors of market efficiency by publicly disseminating the information which is found in the text changes.

[7]For a detailed overview on the covered countries and report languages, please refer to Table 2.

broadly in line with Cohen et al. (2020), despite differences in the holding period, the dataset, and the time horizon. A portfolio based on *bow* yields a significant monthly six-factor alpha of 23 basis points with a t-statistic of 2.3. Put differently, the return of the long-short portfolio is 23 basis points higher than expected. US firm reports thus seem to contain value relevant information that is not timely priced by investors. Using our novel similarity measure *simBERT* instead, we observe a six-factor alpha of 41 basis points with a t-statistic of 4.22. Based on the factor exposures, it seems that the constructed portfolios are rather similar. Even though firms are less likely to amend the standardized SEC 10-K firms without disclosing new information, a long-short portfolio using *simBERT* rather than *bow* yields a 18 basis points larger monthly six-factor alpha in comparison. The difference is significant at the 5% level.

If we apply *simBERT* to non-US annual reports, a portfolio that is long in *non-changers* and short in *changers* achieves a highly significant monthly six-factor alpha of 71 basis points with a t-statistic of 4.95. In contrast, a portfolio based on bow yields a less significant abnormal return of only 29 basis points. The observed 42 basis points performance gain is statistically significant at the 1% level. As in line with our first hypothesis, the difference appears larger for firm reports that lack a standardized report structure. Irrespective of the method applied, we infer that firm reports outside the US also contain value relevant information.

Our results are robust to various dimensions. First, our results are qualitatively reproducible using traditional word-based methods. Second, we show that the results do not suffer from a reversal effect. Third, we find that potential errors in the allocation of publication dates are not likely to explain the results as strategy set-ups with additional investment lags yield comparable alphas.

We further analyze under which circumstances *simBERT* outperforms *bow* by considering the length and language of the firm reports. We find that *simBERT* outperforms *bow* in all tested dimensions. Moreover, we find evidence that the bag-of-words approach does not work well on non-English text. We obtain a negative alpha of 21 basis points which is significant at the 5% level for non-English reports using *bow* in comparison to a significantly positive alpha of 23 basis points using *simBERT*. The difference is statistically significant at the 1% level. A potential explanation could be that the context of words, which is not considered during a word-based approach, might be more important in languages other than English. An alternative explanation is that non-English firm reports might be restructured more often and thus the share of irrelevant changes is higher, leading to a weaker performance of word-based

methods.

As a next step, we try to isolate factors that influence to what extent firms disclose value relevant information. As in line with previous findings (see e.g. Agostino et al., 2011; Devalle et al., 2010; Chalmers et al., 2011), we find evidnce that the reports of firms who disclose according to international accounting standards (IFRS or US-GAAP) contain more value relevant information. Moreover, by separately studying firms from developed and emerging markets, we find that firms who operate in emerging markets disclose substantially less value relevant information. These findings could also be related to our second hypothesis. If firms from emerging markets face less strict regulation and disclosure requirements, we would expect them to disclose less value-relevant information, leading to a smaller abnormal return of our long-short portfolio.

To directly test our second hypothesis, we apply median splits at the country level based on various proxies for securities laws standards (La Porta et al., 2006) and their regulatory environment. We find that the influence of the country's institutional framework is economically substantial. By conducting median splits based on proxies as the liability standard, public enforcement, and the supervisor characteristic, we obtain differences in unexplained monthly returns of up to 74 basis points.

Regarding our third hypothesis, we test whether analysts indeed consider information contained within annual reports. We investigate whether firms that disclose files with less semantically new information receive more positive recommendation and earnings forecast revisions. We find a positive correlation between *simBERT* and the share of positive revisions over all revisions. The effect is observable in and outside the US. We further find evidence that analysts tend to earlier revise their recommendations and forecasts for US stocks. Moreover, it seems that new information in 10-Ks might be more relevant for analysts in comparison to international reports as we observe a higher coefficient for *simBERT* in the US.

To test whether investors actually trade the signals provided by analysts, we exploit the analyst coverage of stocks in the cross-section and obtain highly significant differences in the six-factor alpha of 46 basis points internationally and 43 basis points in the US. If we study smaller and larger stocks separately to disentangle the size effect[8], we observe a similar pattern only in the US. For smaller stocks, we observe a 29 basis point lower alpha for stocks with

[8]Larger stocks tend to be analyzed more often by analysts and thus variables like market capitalization and analyst coverage are highly correlated.

a higher analyst coverage. For larger stocks, the difference in abnormal return is lower with only 16 basis points. Overall, these results provide some evidence that analysts may contribute to market efficiency by conveying qualitative information from accounting statements to the public and signalling it to investors via earnings forecasts and recommendations, a hypothesis that has been supported by Marhfor et al. (2013).

We contribute to the literature in multiple dimensions. First, to the best of our knowledge, we are the first to assess the value relevance of textual information in international firm reports. While Cohen et al. (2020) find that investors overlook value relevant information in US firm reports, we find that value relevant information may also be found in the textual components of international firm reports. Moreover, we find that annual reports of firms that are subject to stricter regulation contain more value-relevant information. Our results are in line with DeFond et al. (2007) who find evidence that earnings announcements contain more informational value in countries with stronger investor protection. More generally, higher disclosure requirements have been associated with a positive effect on stock markets (La Porta et al., 2006) and international capital mobility (Young and Guenther, 2003).

Besides Lang and Stice-Lawrence (2015) who use textual data to investigate the effect of IFRS adoption and top tier auditing on the disclosure quality in a cross-sectional setup, we are not aware of any cross-sectional analysis of the value relevance of textual data in annual reports. A potential reason is that systematically extracting information from documents that lack a standardized structure is not straightforward. Lopez-Lira (2020) constructs risk factors from 10-K files and shows that a textual based factor model achieves an explanatory power that is comparable to traditional factor models. Using financial news, Bybee et al. (2022) derive systematic macroeconomic risk factors by applying topic modelling, latent factor analysis and variable selection. Despite an increase in the usage of textual analysis in finance and accounting, textual disclosure of international firms remains relatively understudied.

Second, we introduce a new framework to detect semantically new information in documents. While word-frequency measures like bag-of-words may be a sufficient proxy for new information in documents with a standardized structure, they may not control for irrelevant document changes. So far, researchers mainly focused on word-based similarity methods, i.e bag-of-words. For example, Cohen et al. (2020) determine the similarity of subsequent US firm reports to test whether investors fully price new textual information contained in firm reports. Next to the above-mentioned study by Lang and Stice-Lawrence (2015), Brown and Knechel (2016)

compare annual reports in the cross-section and find that more similar firms are more likely to hire the same auditor. Some authors also apply bag-of-words to determine the similarity of firms (Hoberg and Phillips, 2016).

The focus on word-based measures is also prevalent in the context of sentiment prediction. Here, most accounting and finance researchers apply the domain-specific dictionaries suggested by Loughran and McDonald (2011) even though fine-tuned sentiment prediction models based on pre-trained language models like BERT promise better accuracy. For example, by considering the context, these sentiment prediction models handle negations by default which is a major advantage. While Araci (2019) suggests *finBERT*, a sentiment prediction model which is fine-tuned on a financial news dataset, Yang et al. (2020) suggest a comparable model which is trained on 10-K and 10-Q files instead. Both models obtain accuracies of more than 86% on the Financial PhraseBank dataset, a publicly available dataset for sentiment prediction within the finance domain. Even though these models have been recognized and applied by some researchers (see e.g. Hiew et al., 2019; Leow et al., 2021), the dictionary approach remains the method of choice for most researchers.

Third, we provide some evidence that US stocks with a higher analyst coverage tend to be more efficiently priced, suggesting that analysts might contribute to market efficiency. Our results add to the accounting literature on the value of sell-side analysts. While Ball and Shivakumar (2008) find that analysts convey new information prior to earnings forecast revisions and thus contribute to market efficiency, Jegadeesh et al. (2004) argue that sell-side analysts tend to recommend so-called *glamour* stocks. These are stocks that are relatively expensive, experience a high growth and volume as well as momentum. As Coleman et al. (2021) show, this tendency is less observable for robo advisors which use NLP methods to extract information from unstructured documents. Robo-advisors are less likely to recommend *glamour* stocks, leading to a long-term outperformance in comparison to stocks that were recommended by human analyst. This observation suggests that digitalization might improve the quality of signals obtained from analysts in the future.

The remainder of this paper is structured as follows. In section 2.1, we introduce our new document similarity measure and compare it to traditional ones. Then, in section 3, we present our international firm report dataset as well as our methodology. We then show our main results in section 4 and discuss various influencing factors in section 5. Finally, we conclude our findings in section 6.

2. simBERT - A semantically sensitive similarity measure

2.1. How SimBERT works

In general, automatically extracting information from international firm reports is challenging. The main reason is that, in contrast to US firms that file their reports in a harmonized HTML format, non-US firms are not bound to an international reporting format. Instead, the majority of firms publish their annual reports in the form of custom-designed PDF files. As a consequence, firm report structures may vary across countries and firms. Besides, non-US firms may also structurally change their reports over time since they are more flexible in comparison to US firms. On the one hand, a firm may choose to exclude or include non-mandatory sections. On the other hand, a company may also choose to rephrase texts from previous firm reports with respect to improving firm reputation rather than providing new information that is relevant for investors. As investors are only interested in semantically new information, we need to identify a similarity measure that is able to filter out irrelevant document changes.

According to Han et al. (2021), similarity measures may be categorized as corpus-, knowledge- and deep learning based. Within the area of textual analysis in finance and accounting, most researchers apply corpus-based methods. Among others, Tetlock (2011), Andreou et al. (2020), Brown and Tucker (2011) and Hoberg and Phillips (2016) apply a *bag-of-words* (*bow*) approach to calculate the textual similarity between documents. While a bag-of-words approach is a suitable method "when researchers intend to compare the exact language between two documents" (Bochkay et al., 2022, p.24), it lacks the possibility to control for irrelevant document changes and thus does not meet our criteria. The same applies for the neural network-based Word2Vec method that, among others, has been applied by Jang et al. (2019).

Knowledge-based methods promise to mitigate this issue by accessing lexical databases like *WordNet* (Miller, 1995) which store semantic relations between words. There are some researchers that use *WordNet* in the context of financial research (see e.g. Hamdan et al., 2013; Hollum et al., 2013). Using knowledge based methods, researchers may potentially identify semantically new information to some extent. However, these databases are monolingual and thus not ideal for our dataset of international, multilingual firm reports. Of course we could translate reports into English, but considering that this process is not only resource consuming, but also might induce a loss of information, depending on the performance of the translation model. Due to these limitations, we refrain from using a knowledge-based similarity measure.

The third group comprises deep learning models. While the concept of training language

models is not new, their performance used to be limited as they often required large labeled datasets that are expensive to create.

However, when Google introduced its pre-trained language model BERT (Bidirectional Transformers for Language Understanding), large labeled datasets were suddenly no longer required. As a transformer model, BERT is capable of learning bidirectional representations from unlabeled text. Instead of processing input from left to right, the model captures the context of words by processing the whole input text at the same time. Within pre-training, the model tries to solve two tasks. First, the model learns to predict masked words within a given text (Masked Language Model). Second, the model aims at predicting whether a specific sentence follows another one (next-sentence prediction). By using a so-called attention mechanism, the model derives weights for the importance of individual words. Put differently, the model identifies important parts of a text and how they connect to other parts.(Devlin et al., 2018) Researchers may fine-tune the pre-trained language model on a small labeled dataset and thereby leverage the general language understanding captured within BERT, which may lead to substantially better results than models which are solely trained on a labeled dataset.

Leveraging the work of Reimers and Gurevych (2020), we propose *simBERT*, a new semantically sensitive document similarity measure. Rather than representing a document on the word-level and thus neglecting contextual information, our measure represents documents on the sentence level. The reason we do not use a pre-trained language model to obtain a vector representation for the entire document is that the multilingual model that we apply may only process a maximum of 128 tokens (words) at a time which is far less than the average amount of words contained within an annual report. We therefore extract the text from annual firm reports and organize it as a list of sentences.[9]. We then obtain state-of-the-art semantically sensitive sentence embeddings by iteratively applying a multilingual sentence-transformer model (Reimers and Gurevych, 2020). For each sentence in a document, we identify the most similar sentence in the previous document by calculating the cosine similarities between the sentence and all sentences in the previous report. We then average the maximum cosine similarities, i. ex. the cosine similarities of all previously identified sentence pairs. A low similarity of two documents, as measured by *simBERT*, indicates that there exist sentences that are not semantically new to any of the sentences of the previous document and thus the document contains semantically new information.

[9] A more detailed description of the text extraction process is provided in section 3.1.

In comparison to other similarity measures, *simBERT* is not symmetric. While sentence B from document D_2 might be the most similar to sentence A from document D_1, the opposite does not necessarily holds true. The reason is that there could be a sentence C in document D_1 that is even more similar to sentence B. This asymmetric attribute can be a strong advantage. For example, if a firm decides to delete a certain part from its annual report while keeping the rest of the document unchanged, it does not disclose any new information. In contrast to a bag-of-words approach, *simBERT* would take a value of 1 and thus correctly indicate that there is no new information in the document.

2.2. A comparison between simBERT and bag-of-words

Let us illustrate the difference of *simBERT* and *bag-of-words* using the 2020 annual report of Anglo American, a mining company headquartered in London. An excerpt of the letter from Stuart Chambers, chairman of the company, is provided in Figure 1. By applying a threshold to the maximum cosine similarities that we obtain during the calculation of our similarity measure, we are able to isolate those sentences that are not sufficiently similar to any of the sentences in the previous report. For example, if the cosine similarity to the most similar sentence is less than 0.75, we might argue that this sentence contains semantically new information. Based on this rule, we highlight sentences that that contain semantically new information.

[Figure 1 about here.]

As we can see, Mr. Chambers mentions the global pandemic and the lockdowns in certain countries "which put additional pressure" on the mining operations of the firm. This information is correctly flagged as new, as he refers to events that happened in the previous year. We observe that the vast majority of sentences is classified as semantically new. While this is not too surprising, given that such letters usually cover events that occured in the previous year, we observe a comparably high share of new information within other sections. Taking this into account, it is surprising though that a *bag-of-words* approach yields a very high cosine similarity of 0.995. We hypothesize that this might be related to the high dimension of the word frequency vector (1x10970). Potentially, high dimensional vectors are more similar by construction and thus the document similarity measure might be biased towards larger reports.

We find evidence that the *bow* approach might indeed be skewed to higher cosine similarities for higher dimensional word frequency vectors. Looking at the one percent of files with the highest and lowest number of unique words, we find that the bow approach yields a cosine

similarity of 0.94 and 0.72 respectively. The difference is highly significant at the 1% level with a t-statistic of 16.75. Gaulin and Peng (2022) observe a similar effect, even though the dimensionality of their term frequency vector is larger by a magnitude of 10. They explain their finding by arguing that for long reports, the probability that sparse word-frequency vectors overlap is higher due to the higher dimensionality.

Comparing the performance of similarity measures with respect to firm reports is not straightforward. Given the size of the documents, hiring humans that manually label the similarity of pairs of subsequent firm reports would be extremely costly. Moreover, even if we sample the assessments of multiple annotators, the derived labels might still be subjective to a certain agree.

To circumvent these problems, we automatically create our own labeled dataset using the Item 1 (business) sections of US annual reports. The dataset is obtained as follows. We subdivide all business sections of US annual reports published in 2021 into two parts, a train and a test sample. Rather than randomly assigning sentences to these two parts, we allocate every fifth (tenth, twentieth) sentence to the training set.[10]

We argue that a powerful similarity measure should indicate the highest similarity for the business section subsets that belong to the same firm. The reason is that it should be the firm itself that conducts the most similar business. Ideally, the accuracy should remain high with a decrease in the size of the training sample and thus fewer information to identify the most similar subset of a business section. To calculate the performance of the different similarity measures, we identify the most similar test sample for a given train sample by calculating the similarity of a training sample with all other test samples. We consider a classification as correct, if the identified pair of training and most similar test sample belong to the business section of the same firm. Finally, we obtain the accuracy of the different similarity measures by relating the number of correct allocations of training and test sample to the total number of allocations.

[Table 1 about here.]

Table 1 shows the accuracy of the different similarity measures based on our labeled dataset of US business sections. If we equally split the business sections into two parts, we obtain an

[10]Note that we only consider business sections of those firms that contain at least 10 (20, 40) sentences to ensure a sufficient amount of data.

accuracy of 65.67% for the bag-of-words approach. A better performance is obtained using our sentence-based similarity measure *simBERT*. Here, we achieve a substantially higher accuracy of 82.00%.

Instead of equally splitting business sections, we also test the performance of the similarity measures if we decrease the size of the training samples. Intuitively, we would expect to see a decline in the performance of the *bag-of-words* approach, given that the probability that a word occurs in both subsections is lower. Indeed, we find that the accuracy drops to 61.97% (41.23% and 19.41%) if we construct the training sample using only every fifth (tenth or twentieth) sentence of the business section. We observe a similar but substantially less pronounced pattern for *simBERT*. Here, the accuracy drops to a minimum of 72.32%, indicating that *simBERT* is substantially more accurate in predicting the semantic similarity of two documents.

3. Data & Methodology

3.1. Data

To analyze the value relevance of international annual reports, we gather annual firm reports for a large amount of international firms taking the following three steps. First, we create lists of stocks[11] for a large number of countries covered by Refinitiv Datastream. We include all developed countries as classified by MSCI except for Japan. The reason for excluding Japan is that Japanese firm reports seem to have a completely different report structure that could not be processed without further domain knowledge. We further collect annual reports for a large amount of firms from emerging countries in accordance with the MSCI classification. As a second step, we iterate over the obtained stock lists and download the corresponding annual PDF reports via Bloomberg. In case more than one annual report was uploaded to Bloomberg within a given year, we choose the latest one available since it likely contains the most accurate information. We download firm reports in the country's main language as long as it is one option of the selection box in Bloomberg.[12] Doing otherwise would risk loosing reports of smaller firms that may not publish their report in more than one language. For all other countries we download English firm reports.[13] Note that we do not consider quarterly

[11] We include delisted stocks to avoid a survivorship bias.

[12] Languages that may be selected within Bloomberg are: English, German, French, Spanish, Italian, Portuguese, Traditional Chinese, Simplified Chinese, Korean.

[13] We provide an extensive overview of the report languages in Table 2.

reports to avoid a too strong bias towards larger firms as quarterly reports are not mandatory in every country and hence mainly available for larger firms.

After collecting the PDF files, we extract the text from electronically readable documents using a Python package *(PDFMiner)*. All other reports are processed using an OCR software *(Abbyy FineReader)*. To ensure that we only use those reports where a sufficient amount of text is available, we exclude reports with less than 100 sentences and file sizes below 10 KB. Next, we split the texts into sentences and remove non-textual information (e.g. line breaks, URLs, and tables).

In addition to extracting text from annual reports, we also determine their publication dates. While it is straightforward to obtain filing dates for US reports from the SEC, determining the publication date of international reports is more challenging. The reason is that the file names of the reports downloaded via Bloomberg only contain the Bloomberg upload date but not the actual publication date. Those dates do not necessarily coincide, in fact we notice batch-wise uploads of firm reports in the early 2000s that cover previous years.

We therefore derive the publication date by extracting all dates mentioned in a report and choosing the one which is closest to but before the Bloomberg upload date as this is very likely the date when the report was finalized. For those cases where we do not identify a date or the derived date coincides with a quarter or fiscal year end, we define the publication date as the minimum of the Bloomberg upload date and the next quarter year end date. This methodology should ensure that we do not base our investments on firm reports that were not actually published then.

Note that for the US, in contrast to all other countries, we collect the data from the SEC using Edgar[14]. We follow the preprocessing steps as described in the internet appendix of Loughran and McDonald (2016) to get rid of the HTML tags included in the SEC file. Afterwards, we apply the same preprocessing steps as described above to obtain a comparable structure.

[Table 2 about here.]

Table 2 summarizes our global dataset of firm reports. In total, we collect reports for 18691 non-US firms across 31 countries. Additionally, we collect more than 180,000 firm reports for

[14]Edgar (https://www.sec.gov/edgar/searchedgar/companysearch.html) is a platform operated by the Security and Exchange Commission in the US and allows the submission and accessing of firm disclosures.

more than 8,000 US firms. We find that the amount of available firm reports differs substantially in the cross-section. Besides the US, we collect the most reports for Australia, India, and the UK, whereas we only download around 300 reports for the Czech Republic. We also allocate countries into *Developed*, *Emerging*, and *Frontier* markets by following the market classification of MSCI. We observe that the average firm report length differs substantially internationally. Firm reports in countries like Brazil on average contain less than three hundred sentences, whereas Spanish firms on average disclose reports with an average length of around 1600 sentences. We also observe a variation in the share of semantically new information across countries. Turkey reports seem to have the lowest share of semantically new information, whereas the most semantically new information may be found in Poland, according to *simBERT*. We further observe differences in the correlation between *simBERT* and *bow*. While English reports seem to have a higher correlation coefficient on average, non-English reports seem to have a lower correlation between these two measures. This might be an indication that non-English reports might contain more irrelevant document changes. Nevertheless, a correlation of around 0.48 and 0.49 in and outside the US indicates that these measures do not measure the same type of document similarity. Our dataset is also diverse with respect to the average market capitalization of a firm. For example, the average firm in Switzerland is valued at $4.17 billion whereas Canada has the lowest average market capitalization of $0.47 billion.

3.2. Detecting value relevant information using factor regressions

In order to test the value relevance of textual data contained in firm reports, we borrow a concept from finance, the calendar time portfolio approach. Generally speaking, we may test whether a certain variable is priced by regressing the return difference of two portfolios of stocks with a high (low) exposure to the variable of interest on well-known pricing factors. If this regression yields a coefficient that is statistically significantly different from zero, we may conclude that the variable of interest carries relevant information that is not adequately priced.

In our context, the variable of interest is the document similarity of two subsequent annual reports. We first sort the entire universe of stocks according to their similarity as measured by the respective measure. We then construct a portfolio that is long (short) in the 20% of stocks with the lowest (*non-changers*) and highest (*changers*) amount of semantically new information. Note that stocks remain in the portfolio for 12 months and are updated on monthly basis, i. ex. newly released reports are added throughout the year. Finally, we regress the return differences on well-known pricing factors. These include the excess market return (Mkt-RF),

the size (SMB) and value (HML) factors of Fama and French (1993), the momentum factor (Carhart, 1997), and the profitability (RMW) and investment (CMA) factors of Fama and French (2015). A significantly positive alpha thus indicates that *changers* underperform. We interpret this underperformance as indication that the reports of *changers* contain (negative) value relevant textual information since the return difference may not be explained by any of the considered factors. Note that we consider different factor models throughout the paper to show that our findings are robust with respect to different model specifications. However, we mainly focus on the six-factor model since it controls for the largest number of risk factors.

4. Value relevance of international annual reports

4.1. US stock market

Within this section we try to replicate the results of Cohen et al. (2020) who find that investors overlook value relevant information in US firm reports. Similar to Cohen et al. (2020), we identify those 20% of firms in a month that change their firm reports the least (most) with respect to the prior year as *non-changers* (*changers*) using different measures of similarity. This allows us to investigate whether our novel approach yields a more precise measure of the value relevance of US firm reports.[15] We then pursue long investments in *non-changers* and short investments in *changers* at the beginning of the month following the publication month. Stocks remain in the portfolio for 12 months. To ensure diversified portfolios, we consider only those months where at least 30 firms were included in each quintile portfolio. As a performance metric, we calculate the alpha obtained from a six-factor model. However, instead of using the market index provided by Fama and French, we use the average stock return of our dataset in a given month as a proxy for market return to control for a potential survivorship bias.[16]

[Table 3 about here.]

Table 3 shows the factor exposure of the different quintile portfolios within the six-factor model. A portfolio that is long in non-changers and short in changers using *bag-of-words* as similarity measure yields a 23 basis point monthly six-factor alpha which is significant at the

[15]Note that in contrast to Cohen et al. (2020), we only consider annual reports (10-K documents) to be consistent with our international dataset of annual reports.

[16]We use the alternative market proxy because we observe significant alphas in all quintiles otherwise. We hypothesize that our dataset might be subject to a survivorship bias induced by missing historical cik-to-cusip mappings of firms who turned bankrupt. Most importantly, the alpha obtained from long-short portfolios does not substantially change if we use our alternative market factor.

5% level. The alpha is driven by both, long and short legs, however the contribution of the long-leg is larger. Our results cannot directly be compared to Cohen et al. (2020), who obtain a three factor alpha of 34 basis points, since we only consider annual reports and use a longer time horizon. For ease of comparison, we restrict our dataset to the same time horizon and extend the dataset by quarterly reports. We obtain a three factor alpha of 40 basis points which is slightly larger than the 34 basis points reported by Cohen et al. (2020). We thus argue that our implementation of the *bag-of-words* approach should be rather similar to theirs.

If we classify *changers* and *non-changers* according to *simBERT*, we find a larger six-factor alpha of 41 basis points which is highly significant. This is a 18 basis point increase in comparison to using the bag-of-words approach as a measure of document similarity. We also test whether this difference is significant by regressing the return differences between the portfolios on the factor exposure[17] and find that is significant at the 5% level and may be traced back to both, long and short legs. This is remarkable, given that the correlation between both measures is substantially higher than in most of the other countries. Moreover, the factor exposure in general is relatively similar to the one obtained from a *bag-of-words* approach. One potential explanation for the performance difference could be that both measures yield highly similar results in most cases but draw substantially different conclusions in some cases.

As long as none of the measures strictly outperforms the other, a combination of these measures may yield even better more precise estimation of similarity. We therefore evaluate a third investment strategy where we construct quintiles based on the average rank of both similarity measures. Given that the 29 basis points monthly six-factor alpha is in between the alphas obtained using the similarity measures individually, *simBERT* indeed seems to outperform the *bag-of-words* approach for US firm reports. Overall we find evidence that US annual reports contain value relevant information.

4.2. International stock markets

In this section, we test whether international annual reports also contain value relevant information. Ideally, we would like to conduct the analysis on the country level. However, since we lack enough reports for some countries, we might not be able to construct sufficiently diversified portfolios. We therefore analyze groups of countries instead. Since US factor data

[17]Note that the factor exposures vary for both portfolios, given that we use country-weighted factors and the portfolios have different country exposures. We therefore regress the return difference on the average of the risk factors.

does not align with international investments, we separately construct country-specific asset pricing factors by following the methodologies described on Kenneth Ronald French's website as closely as possible. To control for a potential survivorship bias, we calculate the market factor by calculating average monthly returns on the country level and the risk-free rate (one-month T-bill rate) from Kenneth French's website. For the remaining factors, we choose a size-decile-based breakpoint of 8 to sort stocks into small-cap and large-cap. All factor returns are measured in U.S. dollars to be consistent with the measurement of the returns of the investigated analyst recommendations and mispricing strategies. To control for country specific effects, we weight country-specific factors in accordance with the portfolio's exposure to a given country.

[Table 4 about here.]

Table 4 shows the factor exposure of the quintile portfolios using *bow*, *simBERT*, and a combination of both as similarity ranking. Using a traditional bag-of-words approach, an equally weighted investments into international stocks outside the US yield a positive alpha of 29 basis points which is statistically significant at the 5% level (t-value of 2.17). In contrast, a long-short portfolios using *simBERT* as similarity measure generates a monthly six-factor alpha of 71 basis point with a t-statistic of 4.95 ($p < 0.001$), indicating that our similarity measure is indeed more precise in measuring document similarity in less harmonized documents. Compared to the *bow* approach, this is a 42 basis point increase in performance on a monthly basis which is significant at the 1% level. This difference in unexplained return for the two similarity measures is also 24 basis points larger in comparison to the US, which is in line with the smaller correlation between both similarity measures presented in section 3.1. Again, we also consider the combination of *simBERT* and *bow* and obtain an alpha of 58 basis points per month. Overall, we interpret our findings as indication that *simBERT* outperforms *bow* with respect to international firm reports.

Compared to the US, we find a 30 basis point increase in unexplained return which is significant at the 1% level. This finding suggests that for non-US reports value relevant information is less timely priced by investors. We will further discuss these hypotheses in section 5.

4.3. Robustness of Results for simBERT

[Figure 2 about here.]

Since we measure value relevance via differences in unexplained returns for *non-changers* and *changers*, we need to show that these differences may not be explained by other stock market phenomena. For example, it could be that investors overestimate the information released in annual reports and thus price them wrongly. Over time, investors would realize their overreaction and correct their mispricing by trading stocks accordingly. We therefore test whether a reversal effect is observable in our setting. We therefore rerun our analysis using different holding periods and plot the cumulated alpha over time. According to Figure 2, we observe no reversal effect for any of the similarity measures. As a consequence, it seems unlikely that the differences in unexplained return are related to an overreaction of investors.

[Table 5 about here.]

Another potential explanation for the observed differences in unexplained returns, at least in international markets, is that the identification of publication dates is wrong in some cases.[18] For example, if we accidentally allocate a date to a report that is prior to its actual publication date, we might base our investment decision on information that is not yet accessible. Thus, we would capture the announcement effect of quantitative information in the report that might positively affect our portfolio returns.

Table 5 addresses the concern that some publication dates might be wrongly identified. We rerun our main calculations using additional lags of one to six months before the start of the investment. We observe comparable alphas if we lag our investments by up to three additional months. For an additional six months delay, the obtained alpha is 36 basis points lower but remains highly statistically significant (2.77). We therefore argue that a potentially wrong allocation of publication dates is not linekly to explain our results.

4.4. Why does simBERT outperform bag-of-words?

While we have shown that our similarity measure seems to outperform *bow* with respect to identifying similar text (see section 2.2), we have not yet investigated under which circumstances

[18] As argued in section 3.1, we have to estimate the publication dates of annual reports as we only know the date when the report was uploaded to Bloomberg.

the outperformance is strongest. For example, we have shown in section 2.1 that the bag-of-words approach may be upwards biased for longer reports. As *simBERT* does not suffer from this effect by construction, it could be that the performance gain is larger for longer reports.

At the same time, the outperformance of *simBERT* might me more pronounced for non-English reports. Even though the bag-of-words approach is not language agnostic, given that word frequency vectors may be constructed for any kind of language, it could be that the context of words is more important in languages other than English. To test whether firm report length or language have an effect on the performance of *simBERT* and *bag-of-words*, we merge the US and international dataset to a global one and apply multiple median splits based on various variables of interest.

[Table 6 about here.]

Table 6 shows the results of various median splits for investments into US and non-US stocks. Primarily, we investigate whether differences to firm report length, as measured by the amount of unique words, or the report language have an impact on the accuracy increase of *simBERT* in comparison to *bow*. Finally, we test whether reports that are filed according to an international accounting standard (IFRS or US-GAAP) contain more value relevant information.

Panel A shows various alphas for investments into international portfolios excluding the US. Most importantly, we observe that *simBERT* outperforms bow in all tested dimensions. For longer reports, we indeed observe a non-significant 31 basis points larger monthly six-factor alpha. In contrast to our initial hypothesis, *simBERT* also outperforms on short firm reports. While *bow* measures a monthly six-factor alpha of 62 basis points for shorter reports, *simBERT* generates a highly significant six-factor alpha of 116 basis points with a t-statistic of 5.65. The difference is highly significant at the 1% level with a t-statistic of 3.81. Thus, it seems that *simBERT* is also a more accurate measure of similarity for shorter reports.

Irrespective of the similarity measure applied, we find that the unexplained return is more pronounced for shorter reports. Using *simBERT* (*bow*), the difference is highly significant at the 1% (5%) level. One potential explanation could be that the probability that new sentences in contain value relevant information is higher. Thus, the difference in unexplained return between investments in *non-changers* and *changers* might be more pronounced here.

simBERT also yields higher alphas for both, English and non-English firm reports. For english reports, we obtain a highly significant alpha of 78 basis points using *simBERT* and 28 basis point lower significant six-factor alpha using *bow*. The difference is significant at the

5% level. For non-English reports, the difference is substantially more pronounced. While we observe a weakly significant alpha of 23 basis points (t-statistic 1.79) for non-English reports using *bow*, we obtain a negative 25 basis point alpha which is significant at the 5% level. We find that the 48 basis point difference is highly significant at the 1% level with a t-statistic of 3.37. This finding suggests that a bag-of-words performs poorly on languages other than English. Potentially, the context of words plays a stronger role in non-English languages. An alternative explanation could be that reports from countries with an official language other than English are restructured more frequently and thus the *bow* measure is distorted.

Overall, we conclude that our similarity measure *simBERT* is a more accurate measure of new information in firm reports. While this particularly holds true in a multilingual setting, researchers may also profit from a more accurate measure of value relevance when conducting their analyses on English reports.

Motivated by the extensive literature on the effects of an implementation of the International Financial Reporting Standards (IFRS) (Chalmers et al., 2011; Karğın, 2013; Mohammadrezaei et al., 2015), we also test whether annual reports of firms that disclose their information using an international accounting standard contain more value relevant information. To do so, we obtain the accounting standards of all firms from Refinitiv. We then split our dataset into two parts by distinguishing based on the employed accounting standard. One group consists of firms that disclose according to an advanced reporting standards like IFRS and US-GAAP, the other one comprises firms who disclose in accordance to other local accounting standards. We obtain results that are in line with the argument of Lang and Stice-Lawrence (2015) that reports of firms which disclose using a more sophisticated accounting scheme contain more value relevant information. We observe a 41 basis point unexplained return for international stocks that file according to an advanced accounting scheme and a weakly-significant 33 basis points abnormal return otherwise.

We also investigate whether the firm report length is correlated with the amount of value relevant information in US annual reports and provide the results in Panel B.[19] For the US, we do not find any significant difference with respect to firm report length. Longer US reports thus are not automatically associated with more value relevant information.

[19]Note that we do not distinguish between language or accounting standard as we do not have any variation within our proxies by construction.

5. Influencing factors

5.1. The Role of the Regulatory environment

As a next step, we try to isolate factors that influence the amount of value relevant information that is disclosed via annual reports. For example, one may argue that firms from more developed countries might disclose more value relevant information. As the regulatory framework might be stronger in developed markets, firms from developed markets might be forced to publish more value relevant information than their peers from emerging markets. To test this hypothesis, we form two groups of countries, developed and emerging market countries using the MSCI market classification.[20]

[Table 7 about here.]

Panel A in Table 7 lists one-, three-, four-, five- and six-factor alphas of investments into stocks from developed and emerging countries. We find evidence that firms who operate in developed markets disclose more value relevant information. We obtain a highly significant monthly six-factor alpha of 74 basis points with a t-statistic of 6.27 which is substantially higher than investments into emerging markets. Here, the obtained alpha is not significantly larger than zero with 46 basis points. These findings are in line with our hypothesis that firms from more developed markets disclose more value relevant information. To test whether the regulatory framework is driving this observation, we first create a regulatory index that combines proxies like *criminal sanctions*, *disclosure requirements*, *investigative powers*, *liability standard*, *orders*, *public enforcement*, *risk making power* and *supervisor characteristics* as proposed by La Porta et al. (2006). We obtain a country ranking for each individual proxy, calculate the mean over all proxies and scale it to a 0-1 range. Thus, the country where firms face the strictest regulatory environment receives the largest value. To ensure that our results are not biased towards the US, we exclude US firm reports from the analysis. Using this regulation index, we construct two subsets from our dataset of international annual reports and estimate the amount of value relevant information disclosed in the reports using the same methodology as before. We observe higher abnormal return in countries that implemented higher regulatory requirements within a one- and three-factor model (see Panel B). While the one-factor (three-factor) alpha is as high as 108 (105) monthly basis points in countries where firms face stricter

[20] Note that we do not consider Frontier markets as we lack enough firm reports to form diversified portfolios for this particular group of countries.

regulation, the alpha is only 55 (50) basis points high within lower regulated countries. If we control for more factors, this observation persists, even though the difference is not significant anymore. These findings further support the argument that a stricter regulatory framework is associated with a larger amount of value relevant information in annual reports.

In Panel C, we separately investigate individual regulatory dimensions and obtain multiple findings. First, we observe the highest difference in value relevance for reports from firms that have a high (low) value along the dimension *supervisor characteristic*. The supervisor characteristic is an index calculated as arithmetic mean of the three dummy variables appointment, tenure and focus. These dummy variables equal one if a majority of the members of the supervisor are not selected by the executive, there is no possibility to dismiss supervisors at the will of the appointing authority and if there is a separate entity taking care of supervising commercial banks (La Porta et al., 2006). First, based on our results, we may argue that firm reports are more informative in countries with more independent supervisors. We observe a highly significant difference of 74 monthly basis point return difference.

Second, considering the difference in abnormal return of portfolios with a high (low) *liability standard*, we may conclude that annual reports of firms from countries with stricter investor protection laws contain more relevant information. The *liability standard* is an index that is calculated by averaging three indices that measure the liability standards for the issuer and its directors, the distributor and the accountants in case of a lawsuit due to misleading statements in the prospectus or audited financial information respectively.(La Porta et al., 2006) For example, a firm that does not change its annual report implies that there are no significant changes that might adversely affect the performance of the firm. At the same time, positive information that remains in the report is more likely to be valid in countries with high investor protection, since failing to disclose negative information would expose the firm to a legal risk. In countries with stricter investor protection, firms are thus more likely to disclose relevant information.

Third, we observe higher levels of value relevance in countries with a more powerful supervision, as indicated by the difference in abnormal return of the stocks from countries with a high (low) *investigative powers* index value. The index *investigative powers* measures the supervisory power to command documents and subpoena testimonies of witnesses in the context of an investigation of a potential violation of securities laws (La Porta et al., 2006). Here, stronger investigative power increases the likelihood that false claims in the annual report may

be identified and thus decrease the probability that a firm willingly files false information.

Finally, we also observe a positive difference of 10 basis points with respect to higher disclosure requirements. This is intuitive, given that firms from countries with higher disclosure requirements will likely include more relevant information in their firm reports, leading to higher alphas. In total, we find strong evidence that annual reports of firms which file in countries with stricter regulation contain more value relevant information.

5.2. The effect of analysts

In the previous sections we have shown that the text of international annual reports contains value relevant information. However, as in line with the findings of Cohen et al. (2020), we observe that investors fail to fully price this information. In principle, there are two potential hypotheses why we observe this abnormal return pattern. First, investors might either be aware of the information contained within firm reports but are unable to trade it due to trading frictions. Second, as automatically extracting information from PDF files is challenging, investors might lack the time to read and process annual reports of the entire universe of firms and thus overlook relevant information.

For the US, Cohen et al. (2020) argue that it is the limited attention of the investors which drives the anomaly. Investors potentially focus on certain (larger) stocks and thus overlook information contained within firm reports of others. A question that arises is whether stock market analysts may contribute to more efficient prices by analyzing information that is contained in firm reports and providing investors with condensed trading signals. This view is supported by the argument that investors "with limited abilities or time to analyze individual securities often rely on the work of sell-side analysts, typically through the analysts' reports" (Bradshaw, 2011, p.2). However, analysts may only contribute to market efficiency if they cover stocks that are not analyzed by investors themselves.

[Figure 3 about here.]

We therefore investigate the distribution of the analyst coverage variable in and outside the US in Figure 3. While the overall shape is similar, we observe a substantially larger amount of stocks with no or only a low analyst coverage within our international dataset[21]. Internationally, around 28% of all stocks are not covered by at least one analyst. For the US,

[21]Note that we assume that there is no analyst covering a stock if we lack analyst data on that particular stock.

this share is substantially lower (11%). While there exist stocks that experience a very high analyst coverage, analysts tend to diversify, especially in the US.

Generally speaking, analyst coverage may only have an impact on on the performance of the strategy if analysts consider textual firm report information within their earnings forecasts or recommendations. Levering the findings of Brown et al. (2015), who show that analysts consider 10-K documents within their earnings forecasts, we hypothesize that analysts consider custom-designed firm reports for international firms in a similar manner. We test this hypothesis by comparing analyst recommendation revisions and earnings forecasts amendments prior and post publication of an annual report. To do so, we differentiate between three different time horizons. On the one hand, we look at revisions of analyst recommendations and earnings forecasts within three months prior and three months post publication of a new firm report. On the other hand, we investigate the time between twelve months and three months before publication as well as the three months post until twelve months post publication. We obtain analyst recommendations and earnings forecasts from the Institutional Brokers' Estimate System (IBES). We measure the analyst reaction to the publication of a firm report using an analyst revision score, which is calculated as follows:

$$ana\_rev_t = \frac{rev\_pos_t - rev\_neg_t}{rev\_pos_t + rev\_neg_t} \tag{1}$$

where rev_pos_t is the number of positive revisions and rev_neg_t the number of negative revisions within a specific month t. We then aggregate the analyst revision score for multiple months and regress it on *simBERT*. Following our argumentation that analysts are able to identify value relevant information from annual reports, *non-changers* should receive more positive revisions than *changers*.

[Table 8 about here.]

Table 8 depicts the regression coefficients of our similarity measure *simBERT*. We find evidence that analysts are more likely to positively revise their one year earnings forecasts for firms with fewer firm report changes between three months prior and post publication date as indicated by the significantly positive coefficient of *simBERT* in column one. A stock whose firm releases a semantically unchanged firm report on average receives 10.63 percentage points more positive next-year earnings revisions in comparison to a firm where the published report contains only semantically new sentences. The effect is even stronger in the US. Here, we

obtain a highly significant coefficient of 36.53 suggesting that non-changers receive up to 36.53 percentage points more positive earnings revisions within three months before and three months after the publication of the annual report.

We further find evidence that analysts react to new information being released within twelve and three months before the publication of the firm report. This effect is slightly stronger in the US which might indicate that analysts react faster to new information being released prior to the publication of the firm report than outside the US. We obtain a significant coefficient of 37.69. Outside the US, the opposite is the case. This might indicate that in non-US markets, less firm report information is publicly known prior to its publication. Given that the level of document similarity does not affect the earnings revision score during three and twelve months after the publication of the firm report, we argue that the effects we measure should indeed be related to information that is discussed within annual reports.

Our results are also robust with respect to longer-term predictions. For the two-year fiscal earnings forecasts, we recognize the same effects as before. We obtain an even higher significant global coefficient of 8.94 with a t-statistic of 2.43 between three months before and after the release of a firm report. Again, we observe stronger reactions in the US both around the publication date and before, indicating that analysts might consider changes to 10-Ks as more relevant than changes to custom-designed annual reports.

If we consider revisions to analyst recommendations, we detect a similar but less pronounced effect. Firms that file reports with less semantically new information receive on average more positive revisions than firms who disclose more semantically new information. This effect is primarily driven by non-US firms.

If investors ultimately trade these signals, we should see more efficient prices for stocks with a higher analyst coverage. We therefore investigate whether long-short investments into stocks with a higher analyst coverage generate smaller alpha.

[Table 9 about here.]

Table 9 shows the monthly alphas obtained from *simBERT* driven long-short investments into different subsamples of our non-US and US datasets. Internationally, we observe that *simBERT* driven long-short investments into stocks with a higher analyst coverage generate a 46 basis point lower six-factor alpha than investments into stocks with a lower analyst coverage. The difference is significant at the 1% level. Within the US, we observe a similar pattern. Here, the difference is even larger with 53 basis points and significant at the 1% level. However, analyst

coverage might not be the only driver behind the difference in return. The reason is that firms with a higher analyst coverage tend to be larger. Any differences between those two subgroups could thus also be related to the size effect which predicts more efficient prices for larger stocks due to fewer trading frictions. We therefore combine a median split based on size with the median split based on analyst coverage to disentangle the analyst effect from the size effect.

Controlling for size, we find that an equally weighted long-short investment into smaller international stocks that have a lower analyst coverage yield a highly significant alpha of 83 basis points with a t-statistic of 3.92. Investments into smaller stocks that have a higher analyst coverage yield a slightly lower monthly alpha of 77 basis points. For larger international stocks, we observe a similar effect. We thus conclude that the difference in unexplained return is mostly explained by firm size internationally.

Considering investments into US stocks, we observe a difference in the six-factor alphas for smaller and larger stocks. For smaller stocks, the difference is 29 basis points, whereas for larger stocks it is 16 basis points. Even though that these differences are not significant due to the higher return variation in the smaller sized portfolios, we may interpret these findings as indication that analysts might contribute to more efficient prices by conveying the textual information to the pblic through earnings and recommendation revisions.

6. Conclusion

In this paper, we investigate whether international annual reports contain value relevant information. To do so, we construct long-short portfolios based on international stocks whose reports contain the least (most) amount of semantically new information. We then regress the portfolio returns on well-known pricing factors. If this regression yields a coefficient that is statistically significantly different from zero, we may conclude that annual reports contain value relevant information that is not timely priced by investors.

To measure the share of semantically new information in firm reports, we introduce *simBERT*, a new similarity measure that leverages latest advances in NLP from computer science. Doing so, we address recent calls to employ new approaches in textual analysis that rely on machine learning, and in particular deep learning (Bochkay et al., 2022). By examining the semantic similarity of sentences, we show that *simBERT* is a superior measure of document similarity in comparison to the traditional bag-of-words due to its ability to filter out irrelevant document changes. Moreover, researchers do not have to deal with preprocessing steps like stop-

word removal, lemmatization or stemming. As the pre-trained models are publicly available, the process should also be more transparent and easier to replicate for other researchers.

Similar to the results Cohen et al. (2020) obtain for the US market, we find that international annual reports contain value relevant information. Specifically, a strategy which is long in *non-changers* and short in *changers* yields an economically large, statistically significant monthly six-factor alpha of 71 basis points. Moreover, while the *simBERT* strategy outperforms the *bag-of-words* approach applied by Cohen et al. (2020) on US and international firm reports, the spread is stronger internationally. These findings are consistent with the idea that irrelevant document changes are more likely to occur in documents that lack a fixed report structure.

Among international markets, we find less value relevant information in emerging countries, for which we obtain a non-significant six-factor alpha of 46 basis points per month. To the extent that the *non-changers*-minus-*changers* strategy loads up on value-relevant information that is communicated in firm reports, its effectiveness will depend on the quality of financial statements and reporting requirements, which could explain the lower performance for emerging markets. Using security laws proxies of La Porta et al. (2006), we find strong evidence that the value relevance is indeed larger in countries with stricter regulatory requirements.

We further investigate if financial analysts may contribute to market efficiency by conveying value-relevant information that can be found in the changes of financial statements to the public. Overall, our analyses support this hypothesis. First, in the months surrounding the financial statement publication, analysts are more likely to downgrade (upgrade) their earnings forecasts and recommendations for stocks with low (high) document similarly based on *simBERT*. Second, we find that the strategy tends to generate a higher (lower) performance for firms with low (high) analyst coverage.

While we demonstrate the power of textual analysis with *BERT* to determine the value relevance of firm report information, our similarity measure can be applied to other tasks in finance and accounting. Potential examples include competitor identification (see e.g. Hoberg and Phillips, 2016), topic modelling, portfolio selection, or improved automated stock recommendations. For instance, in a recent study, Cao et al. (2021) show that an AI-driven analyst may already outperform human analysts for firms that disclose transparent and large amounts of information. With an ever-growing amount of textual data and more powerful language models, it might be a matter of time until artificial analysts also outperform humans in more difficult settings.

References

Agostino, M., Drago, D., and Silipo, D. B. (2011). The value relevance of ifrs in the european banking industry. *Review of quantitative finance and accounting*, 36(3):437–457.

Andreou, P. C., Harris, T., and Philip, D. (2020). Measuring firms' market orientation using textual analysis of 10-k filings. *British Journal of Management*, 31(4):872–895.

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Azevedo, V. and Müller, S. (2021). Analyst recommendations and mispricing across the globe. *Available at SSRN 3705141*.

Ball, R. and Shivakumar, L. (2008). How much new information is there in earnings? *Journal of Accounting Research*, 46(5):975–1016.

Barth, M. E., Landsman, W. R., and Lang, M. H. (2008). International accounting standards and accounting quality. *Journal of accounting research*, 46(3):467–498.

Bochkay, K., Brown, S. V., Leone, A. J., and Tucker, J. W. (2022). Textual analysis in accounting: What's next? *Available at SSRN*.

Bradshaw, M. T. (2011). Analysts' forecasts: what do we know after decades of work? *Available at SSRN 1880339*.

Brown, L. D., Call, A. C., Clement, M. B., and Sharp, N. Y. (2015). Inside the "black box" of sell-side financial analysts. *Journal of Accounting Research*, 53(1):1–47.

Brown, S. V. and Knechel, W. R. (2016). Auditor–client compatibility and audit firm selection. *Journal of Accounting Research*, 54(3):725–775.

Brown, S. V. and Tucker, J. W. (2011). Large-sample evidence on firms' year-over-year md&a modifications. *Journal of Accounting Research*, 49(2):309–346.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Bybee, L., Kelly, B. T., and Su, Y. (2022). Narrative asset pricing: Interpretable systematic risk factors from news text. *Johns Hopkins Carey Business School Research Paper*, (21-09).

Cao, S., Jiang, W., Wang, J. L., and Yang, B. (2021). From man vs. machine to man + machine: The art and ai of stock analyses. Technical report, National Bureau of Economic Research.

Carhart, M. M. (1997). On persistence in mutual fund performance. *Journal of Finance*, 52(1):57–82.

Chalmers, K., Clinch, G., and Godfrey, J. M. (2011). Changes in value relevance of accounting information upon ifrs adoption: Evidence from australia. *Australian journal of management*, 36(2):151–173.

Chandrasekaran, D. and Mago, V. (2021). Evolution of semantic similarity—a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–37.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. (2022). Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Cohen, L., Malloy, C., and Nguyen, Q. (2020). Lazy prices. *The Journal of Finance*, 75(3):1371–1415.

Coleman, B., Merkley, K. J., and Pacelli, J. (2021). Human versus machine: A comparison of robo-analyst and traditional research analyst investment recommendations. *The Accounting Review, Forthcoming*.

DeFond, M., Hung, M., and Trezevant, R. (2007). Investor protection and the information content of annual earnings announcements: International evidence. *Journal of Accounting and Economics*, 43(1):37–67.

Devalle, A., Onali, E., and Magarini, R. (2010). Assessing the value relevance of accounting data after the introduction of ifrs in europe. *Journal of international financial management & accounting*, 21(2):85–119.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of financial economics*, 116(1):1–22.

Gaulin, M. and Peng, X. (2022). Peering into compensation disclosure: Semantic similarity and peer selection.

Guo, L., Li, F. W., and Wei, K. J. (2020). Security analysts and capital market anomalies. *Journal of Financial Economics*, 137(1):204–230.

Hamdan, H., Béchet, F., and Bellot, P. (2013). Experiments with dbpedia, wordnet and sentiwordnet as resources for sentiment analysis in micro-blogging. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 455–459.

Han, M., Zhang, X., Yuan, X., Jiang, J., Yun, W., and Gao, C. (2021). A survey on the techniques, applications, and performance of short text semantic similarity. *Concurrency and Computation: Practice and Experience*, 33(5):e5971.

Hiew, J. Z. G., Huang, X., Mou, H., Li, D., Wu, Q., and Xu, Y. (2019). Bert-based financial sentiment index and lstm-based stock return predictability. *arXiv preprint arXiv:1906.09024*.

Hoberg, G. and Phillips, G. (2016). Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, 124(5):1423–1465.

Hollum, A. T. G., Mosch, B. P., and Szlávik, Z. (2013). Economic sentiment: Text-based prediction of stock price movements with machine learning and wordnet. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 322–331. Springer.

Jang, B., Kim, I., and Kim, J. W. (2019). Word2vec convolutional neural networks for classification of news articles and tweets. *PloS one*, 14(8):e0220976.

Jegadeesh, N., Kim, J., Krische, S. D., and Lee, C. M. (2004). Analyzing the analysts: When do recommendations add value? *The journal of finance*, 59(3):1083–1124.

Karğın, S. (2013). The impact of ifrs on the value relevance of accounting information: Evidence from turkish firms. *International Journal of Economics and Finance*, 5(4):71–80.

La Porta, R., Lopez-de Silanes, F., and Shleifer, A. (2006). What works in securities laws? *The journal of finance*, 61(1):1–32.

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

Leow, E. K. W., Nguyen, B. P., and Chua, M. C. H. (2021). Robo-advisor using genetic algorithm and bert sentiments from tweets for hybrid portfolio optimisation. *Expert Systems with Applications*, 179:115060.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lopez-Lira, A. (2020). Risk factors that matter: Textual analysis of risk disclosures for the cross-section of returns. *Jacobs Levy Equity Management Center for Quantitative Financial Research Paper*.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, 66(1):35–65.

Loughran, T. and McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4):1187–1230.

Marhfor, A., M'Zali, B., Cosset, J.-C., and Charest, G. (2013). Stock price informativeness and analyst coverage. *Canadian Journal of Administrative Sciences/Revue Canadienne des Sciences de l'Administration*, 30(3):173–188.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Mohammadrezaei, F., Mohd-Saleh, N., and Banimahd, B. (2015). The effects of mandatory ifrs adoption: A review of evidence based on accounting standard setting criteria. *International Journal of Disclosure and Governance*, 12(1):29–77.

Reimers, N. and Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Tetlock, P. C. (2011). All the news that's fit to reprint: Do investors react to stale information? *The Review of Financial Studies*, 24(5):1481–1512.

Yang, Y., Uy, M. C. S., and Huang, A. (2020). Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*.

Young, D. and Guenther, D. A. (2003). Financial reporting environments and international capital mobility. *Journal of Accounting Research*, 41(3):553–579.

Figure 1: Excerpt from Anglo American's 2020 Annual Report. Highlighted text contains semantically new information. We define a sentence as containing semantically new information if it is not sufficiently similar to at least one sentence in the previous report. We therefore compare the sentence embeddings obtained from a sentence-transformer model by calculating pairwise cosine similarities. We define a sentence pair as sufficiently similar if the cosine similarity is at least as high as 0.75.

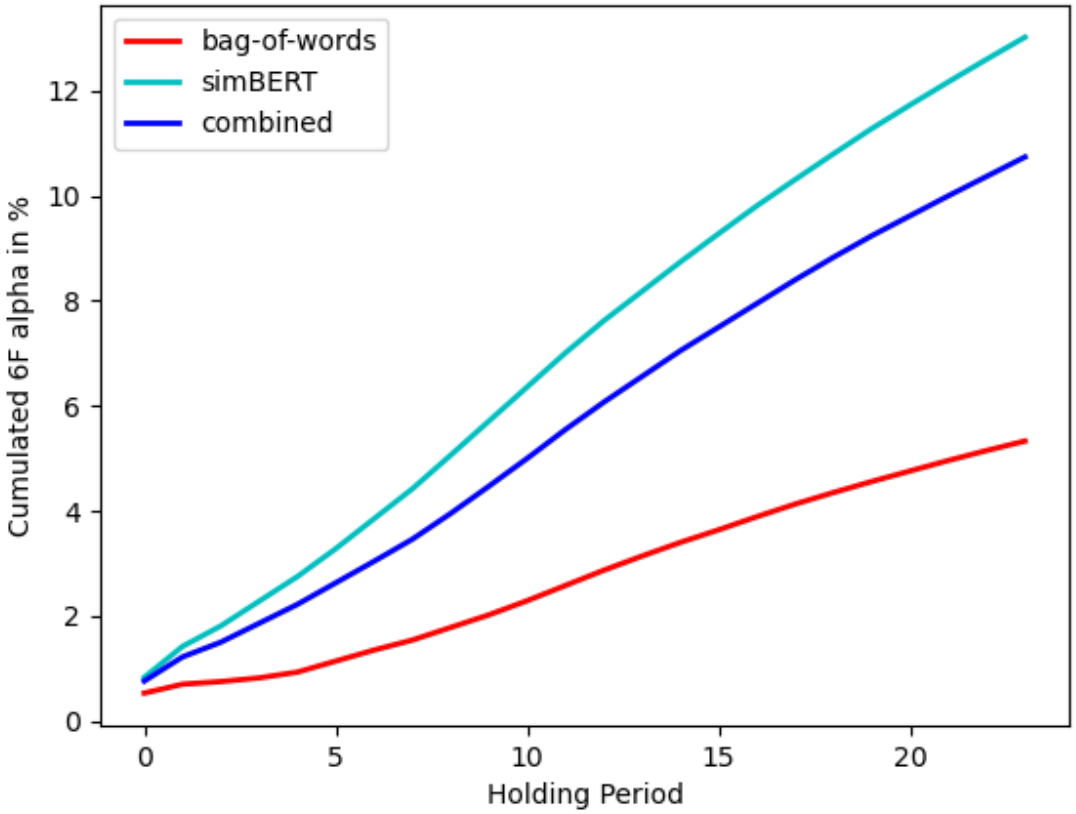

Figure 2: We plot the cumulated six-factor alpha of portfolios that are long in *non-changers* and short in *changers* based on different similarity measures and and different monthly holding periods.

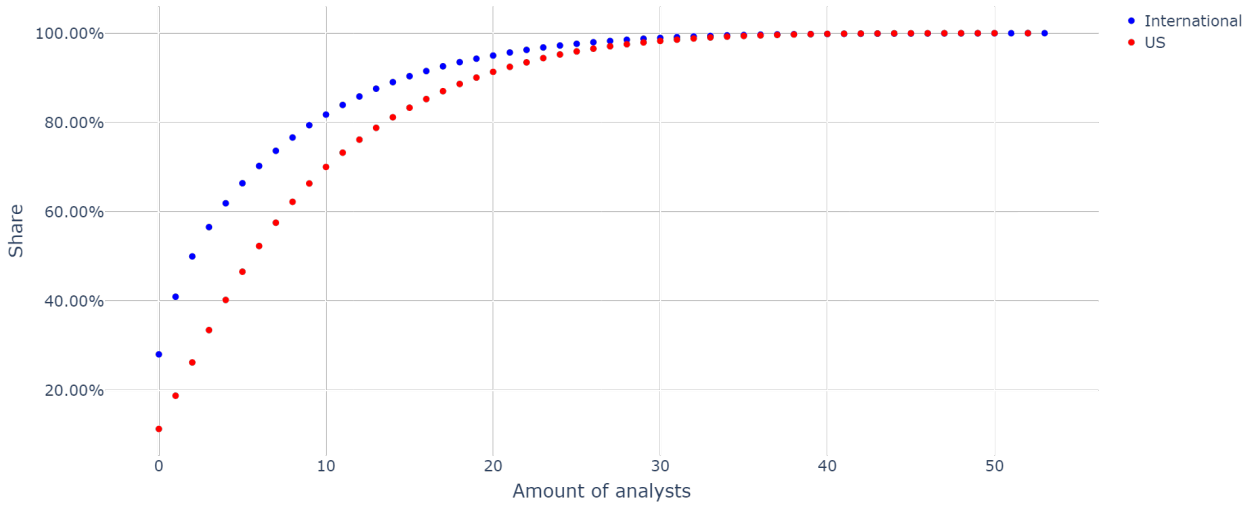

Figure 3: Cumulated distribution of analyst coverage within the US and internationally.

Table 1: **Performance of different similarity measures**

Similarity measure	(1)	(2)	(3)	(4)
bag-of-words	65.67	61.97	41.23	19.41
simBERT	82.00	78.24	75.58	72.32
#Firms	4790	4776	4738	4633

This table provides an overview of the accuracy of different similarity measures. We measure the accuracy by relating the amount of correct allocations of two subsets of the business section to the amount of total allocations. Column (1) shows the results for a 1:1 relation of the *training* and *test* set. We calculate the similarity of an observation in the training set, which is the concatenation of every second sentence of a firm's business section, with all observations in the test set. An observation in the test set comprises all other sentences of a firm's business section. Column (2), column (3) and column (4) show the results for 1:4, 1:9 and 1:19 splits respectively.

Table 2: **Report Statistics by Country**

Country	Lang.	MSCI	#firms	#Reports	Sentences	simBERT	BOW	Corr.	MV
Ex-US*			18691	323741	674	0.79	0.93	0.48	1.50
Australia	English	Dev.	2673	52434	552	0.80	0.95	0.59	1.03
Austria	German	Dev.	148	2220	901	0.74	0.92	0.46	0.91
Belgium	English	Dev.	125	1893	931	0.77	0.90	0.46	2.39
Brazil	Port.	Emer.	204	3686	277	0.81	0.95	0.38	3.08
Canada	English	Dev.	1082	25030	296	0.84	0.96	0.47	0.47
Chile	Spanish	Emer.	204	2861	346	0.76	0.98	0.46	0.96
Czech Rep	English	Emer.	32	325	1162	0.76	0.86	0.59	1.63
Denmark	English	Dev.	175	1807	551	0.76	0.94	0.62	1.46
Finland	English	Dev.	195	3853	634	0.75	0.94	0.47	1.73
France	French	Dev.	1131	17085	598	0.78	0.89	0.44	2.48
Germany	German	Dev.	1225	16426	1056	0.82	0.93	0.49	2.21
Greece	English	Emer.	142	1656	541	0.77	0.93	0.51	0.66
India	English	Emer.	2965	52449	834	0.81	0.93	0.39	0.66
Ireland	English	Dev.	105	1705	816	0.83	0.98	0.59	1.04
Italy	Italian	Dev.	386	7235	1457	0.84	0.95	0.52	1.46
Mexico	Spanish	Emer.	79	785	728	0.79	0.92	0.17	2.94
Netherlands	English	Dev.	167	1785	621	0.76	0.59	0.64	2.68
New Zealand	English	Dev.	230	4884	452	0.69	0.83	0.62	0.49
Norway	English	Dev.	96	1176	646	0.75	0.93	0.47	0.60
Peru	Spanish	Emer.	147	1647	242	0.82	0.93	0.41	0.65
Poland	English	Emer.	118	1430	629	0.73	0.93	0.46	1.00
Portugal	Port.	Dev.	115	1637	725	0.79	0.94	0.42	0.93
Russia	Russian	Frontier	363	6343	607	0.78	0.83	0.64	2.90
Singapore	English	Dev.	921	16250	614	0.74	0.91	0.57	0.64
South Africa	English	Emer.	728	11549	602	0.74	0.94	0.46	1.10
Spain	Spanish	Dev.	317	6037	1618	0.82	0.94	0.50	3.06
Sweden	English	Dev.	530	6910	738	0.75	0.94	0.58	1.49
Switzerland	German	Dev.	233	3283	1321	0.86	0.93	0.41	4.17
Turkey	Turkish	Emer.	411	5225	298	0.90	0.94	0.59	0.54
UK	English	Dev.	3344	43283	623	0.79	0.96	0.53	1.87
USA	English	Dev.	8205	185694	959	0.86	0.98	0.47	3.21

This Table provides a broad overview on our international firm report dataset. We state the report language on the country level and report the MSCI classification for each country. We further report the average number of firms and reports per country as well as the average number of sentences within a report. *simBERT* and *BOW* show the average similarity scores of our newly proposed and a traditional bag-of-words similarity measure. The correlation column yields the correlation of *simBERT* and *bag-of-words*. We further report the average market capitalization of a firm within a country.

Table 3: **Calendar-time portfolio returns: US**

Sim. Meas.	factor	Q1	Q2	Q3	Q4	Q5	Q1-Q5
	MKTRF	1.0***	1.0***	1.02***	1.01***	0.98***	0.02
		(57.85)	(59.01)	(69.79)	(64.15)	(41.71)	(0.49)
	SMB	0.02	0.01	-0.01	0.01	-0.01	0.02
		(0.61)	(0.39)	(-0.32)	(0.35)	(-0.17)	(0.46)
	HML	0.01	0.06**	-0.0	-0.07***	0.0	0.0
		(0.23)	(2.31)	(-0.09)	(-2.53)	(0.11)	(0.03)
Bag-of-words	RMW	0.09***	-0.01	-0.09***	-0.05*	0.02	0.07
		(2.48)	(-0.25)	(-3.26)	(-1.91)	(0.59)	(1.12)
	CMA	0.03	0.03	0.0	0.02	-0.07	0.11
		(0.68)	(0.76)	(0.07)	(0.66)	(-1.35)	(1.22)
	WML	0.04**	0.01	0.01	-0.02	-0.01	0.05**
		(1.95)	(0.48)	(0.24)	(-0.98)	(-0.77)	(1.94)
	Alpha	**0.12****	**0.04**	**0.09***	**0.0**	**-0.1***	**0.23****
		(2.03)	**(0.89)**	**(1.66)**	**(-0.04)**	**(-1.74)**	**(2.3)**
	MKTRF	0.9***	1.0***	1.02***	1.06***	1.03***	-0.13***
		(38.5)	(48.05)	(82.32)	(72.52)	(52.98)	(-4.3)
	SMB	0.07**	0.03	-0.05**	-0.06**	0.02	0.05
		(1.96)	(1.31)	(-2.29)	(-1.94)	(0.66)	(0.98)
	HML	0.07**	-0.02	0.01	0.03	-0.1***	0.17***
		(2.41)	(-0.57)	(0.27)	(0.95)	(-2.46)	(3.11)
simBERT	RMW	0.08*	0.08**	-0.02	-0.05	-0.13***	0.21***
		(1.91)	(2.3)	(-0.72)	(-1.55)	(-4.28)	(4.03)
	CMA	0.02	0.08**	0.07**	-0.09**	-0.05	0.07
		(0.4)	(1.97)	(2.03)	(-1.97)	(-1.22)	(1.05)
	WML	0.06***	0.03*	0.0	-0.01	-0.06**	0.12***
		(2.83)	(1.73)	(0.02)	(-0.73)	(-2.34)	(3.0)
	Alpha	**0.24*****	**0.1***	**0.06**	**-0.07**	**-0.17*****	**0.41*****
		(4.16)	**(1.89)**	**(1.09)**	**(-1.28)**	**(-2.82)**	**(4.22)**
	MKTRF	0.94***	1.02***	1.0***	1.03***	1.02***	-0.09***
		(52.15)	(71.86)	(61.98)	(70.47)	(50.63)	(-2.8)
	SMB	0.07**	-0.01	-0.0	-0.02	-0.01	0.08
		(2.18)	(-0.66)	(-0.0)	(-0.7)	(-0.44)	(1.6)
	HML	0.06**	0.0	-0.01	-0.01	-0.05	0.11*
		(2.27)	(0.04)	(-0.27)	(-0.33)	(-1.16)	(1.85)
Combined	RMW	0.09***	0.06**	-0.06**	-0.09***	-0.04	0.14**
		(2.84)	(2.1)	(-2.11)	(-3.14)	(-1.16)	(2.29)
	CMA	0.01	0.07*	0.1***	-0.05	-0.11**	0.12
		(0.17)	(1.86)	(2.76)	(-1.49)	(-2.33)	(1.44)
	WML	0.05**	0.04***	-0.01	-0.01	-0.05**	0.1***
		(2.37)	(3.31)	(-0.56)	(-0.71)	(-2.15)	(2.47)
	Alpha	**0.17*****	**0.11****	**0.1****	**-0.09***	**-0.12***	**0.29*****
		(2.99)	**(2.12)**	**(2.04)**	**(-1.72)**	**(-1.86)**	**(2.76)**

US calendar-time portfolio regression based on *simBERT*, *bag-of-words* and an equally weighted combination of both measures. Quintile 1 contains firms with the most similar reports. Stocks enter the portfolio one month after their publication and remain in the portfolio for 12 months. We consider investments from January 1998 until June 2021. We report quintile exposures with respect to a Fama French six-factor model. We denote the t-statistics of the coefficients in parentheses. * indicates significance at the 10% level, ** indicates significance at the 5% level and *** indicates significance at the 1% level.

Table 4: **Calendar-time portfolio returns: International stocks outside the US**

	factor	$Q1_{sB}$	$Q2_{sB}$	$Q3_{sB}$	$Q4_{sB}$	$Q5_{sB}$	$Q1\text{-}Q5_{sB}$
	MKTRF	0.99***	1.05***	1.07***	1.09***	1.08***	-0.01
		(77.13)	(124.54)	(64.67)	(90.43)	(90.0)	(-0.48)
	SMB	-0.01	0.03	0.01	0.02	-0.08***	0.04
		(-0.36)	(0.98)	(0.29)	(0.6)	(-3.8)	(0.62)
	HML	0.12***	0.01	0.02	-0.12***	-0.08***	0.11
		(4.15)	(0.64)	(0.62)	(-4.78)	(-2.94)	(1.58)
Bag-of-words	RMW	0.01	-0.02	-0.06	-0.01	0.0	0.02
		(0.26)	(-0.72)	(-1.09)	(-0.36)	(0.01)	(0.27)
	CMA	-0.05	0.01	-0.08	0.06*	-0.01	0.05
		(-1.58)	(0.29)	(-0.84)	(1.69)	(-0.32)	(0.51)
	WML	-0.01	-0.02	0.04*	-0.03**	-0.02	-0.01
		(-0.69)	(-1.13)	(1.88)	(-1.95)	(-1.37)	(-0.43)
	alpha	**0.27*** **	**0.18*** **	**0.01**	**0.03**	**-0.02**	**0.29** **
		(5.38)	**(3.42)**	**(0.15)**	**(0.6)**	**(-0.32)**	**(2.17)**
	MKTRF	0.99***	1.04***	1.05***	1.11***	1.09***	-0.02
		(51.07)	(84.49)	(87.92)	(103.61)	(78.01)	(-0.5)
	SMB	0.14***	0.03	0.02	-0.07***	-0.09**	0.27***
		(3.26)	(1.27)	(0.92)	(-2.97)	(-2.38)	(2.95)
	HML	0.08*	0.08***	0.01	-0.1***	-0.14***	0.22***
		(1.71)	(2.87)	(0.54)	(-4.34)	(-4.61)	(3.37)
simBERT	RMW	-0.0	-0.0	0.07*	0.01	-0.04	0.09
		(-0.1)	(-0.01)	(1.92)	(0.17)	(-1.55)	(1.23)
	CMA	-0.06	0.01	0.06	0.05	-0.04	-0.02
		(-1.0)	(0.2)	(1.22)	(0.89)	(-0.99)	(-0.21)
	WML	0.05***	-0.0	-0.05***	-0.04***	-0.01	0.05
		(2.59)	(-0.07)	(-2.77)	(-2.77)	(-0.75)	(1.24)
	Alpha	**0.58*** **	**0.15*** **	**0.06**	**-0.14** **	**-0.08**	**0.71*** **
		(7.37)	**(2.34)**	**(-1.14)**	**(-2.04)**	**(-1.32)**	**(4.95)**
	MKTRF	0.99***	1.05***	1.04***	1.1***	1.1***	-0.02
		(79.67)	(96.66)	(71.88)	(97.17)	(72.9)	(-0.66)
	SMB	0.07**	0.02	0.05*	-0.05**	-0.09***	0.18**
		(2.37)	(0.85)	(1.66)	(-1.92)	(-3.29)	(2.25)
	HML	0.1***	0.07***	0.02	-0.11***	-0.14***	0.19***
		(3.18)	(2.79)	(1.01)	(-4.37)	(-4.6)	(2.85)
Combined	RMW	0.0	0.02	-0.01	-0.02	-0.02	0.07
		(0.13)	(0.61)	(-0.41)	(-0.76)	(-0.95)	(1.0)
	CMA	-0.09**	0.08**	-0.01	0.06	-0.01	-0.03
		(-2.09)	(2.28)	(-0.34)	(1.12)	(-0.27)	(-0.36)
	WML	0.02	-0.02	-0.01	-0.03**	-0.01	0.02
		(1.62)	(-1.5)	(-0.58)	(-2.43)	(-0.75)	(0.5)
	alpha	0.49***	0.12**	-0.01	-0.09	-0.07	0.58***
		(8.88)	(2.36)	(-0.22)	(-1.61)	(-1.19)	(4.27)
	Alpha	**0.49*** **	**0.12** **	**-0.01**	**-0.09**	**-0.07**	**0.58*** **
		(8.88)	**(2.36)**	**(-0.22)**	**(-1.61)**	**(-1.19)**	**(4.27)**

We run international calendar-time portfolio regressions based on *simBERT*, *bag-of-words* and an equally weighted combination of both measures. The US is excluded from the analysis. We report quintile exposures with respect to a six-factor model using international factor data. Quintile 1 contains firms with the most similar reports. Stocks enter the portfolio one month after their publication and remain in the portfolio for 12 months. We consider investments from January 1998 until June 2021. We denote the t-statistics of the coefficients in parentheses. * indicates significance at the 10% level, ** indicates significance at the 5% level and *** indicates significance at the 1% level.

Table 5: **Robustness test: Lagged investments**

	Global*			USA		
Similarity measure	1F	3F	6F	1F	3F	6F
simBERT	0.92***	0.87***	0.71***	0.61***	0.63***	0.41***
	(8.31)	(8.06)	(4.95)	(5.25)	(6.55)	(4.22)
simBERT_l1	0.9***	0.85***	0.67***	0.58***	0.61***	0.39***
	(8.05)	(7.9)	(4.58)	(5.16)	(6.45)	(4.2)
simBERT_l2	0.84***	0.79***	0.56***	0.55***	0.58***	0.38***
	(6.87)	(6.87)	(3.48)	(5.02)	(6.32)	(4.2)
simBERT_l3	0.92***	0.87***	0.65***	0.51***	0.55***	0.35***
	(7.08)	(7.17)	(3.86)	(4.82)	(6.1)	(4.02)
simBERT_l6	0.78***	0.75***	0.45***	0.45***	0.48***	0.3***
	(6.34)	(6.06)	(2.77)	(4.05)	(4.97)	(3.04)

*Investments into thirty countries outside the US.
This table shows different factor exposures of lagged calendar-time portfolio regressions based on *simBERT*-related for non-US and US stocks separately. We denote the t-statistics of the coefficients in parentheses. * indicates significance at the 10% level, ** indicates significance at the 5% level and *** indicates significance at the 1% level.

Table 6: **simBERT vs. bag-of-words: Performance comparison in various dimensions**

	simBERT			bag-of-words		
	1F	3F	6F	1F	3F	6F
Panel A: Global (Ex-US)						
Total	0.92***	0.87***	0.71***	0.34***	0.29***	0.29**
	(8.31)	(8.06)	(4.95)	(3.32)	(2.81)	(2.17)
Short report	1.37***	1.37***	1.16***	0.58***	0.59***	0.62***
	(8.61)	(8.45)	(5.65)	(4.08)	(3.94)	(3.13)
Long report	0.74***	0.69***	0.55***	0.37***	0.33***	0.24
	(6.08)	(5.61)	(3.33)	(2.93)	(2.58)	(1.62)
English	0.98***	0.96***	0.78***	0.44***	0.36***	0.5***
	(8.25)	(7.48)	(4.97)	(4.03)	(3.29)	(3.63)
Non-English	0.41***	0.34***	0.23*	-0.18*	-0.19*	-0.21**
	(3.18)	(2.95)	(1.79)	(-1.78)	(-1.91)	(-1.96)
IFRS/US-GAAP	0.43***	0.42***	0.41***	0.34***	0.29***	0.29**
	(3.73)	(3.61)	(2.8)	(3.32)	(2.81)	(2.17)
Local	0.48***	0.46***	0.33*	0.17	0.13	0.04
	(3.87)	(3.73)	(1.9)	(1.52)	(1.17)	(0.29)
Panel B: US						
Total	0.61***	0.63***	0.41***	0.32***	0.33***	0.23**
	(5.25)	(6.55)	(4.22)	(3.49)	(3.66)	(2.3)
Short report	0.57***	0.59***	0.4***	0.25**	0.25**	0.15
	(4.53)	(5.08)	(3.27)	(2.24)	(2.24)	(1.22)
Long report	0.57***	0.6***	0.43***	0.33***	0.34***	0.33***
	(4.61)	(5.54)	(3.74)	(2.67)	(2.83)	(2.48)

Global calendar-time portfolio regressions based on *simBERT* and *bag-of-words*. We report the long-short one-, three, and six-factor alphas of investments into various portfolios that are constructed among dimensions like firm report length and language as well as accounting standard. Note that we apply a median split to differentiate short from long firm reports. Stocks enter the portfolio one month after their publication and remain in the portfolio for 12 months. We consider investments from January 1998 until June 2021. We denote the t-statistics of the coefficients in parentheses. * indicates significance at the 10% level, ** indicates significance at the 5% level and *** indicates significance at the 1% level.

Table 7: **Calendar-time portfolio returns: Median splits based on market development and regulatory environment**

Name	Subgroup	1F	3F	4F	5F	6F
Panel A: Development Status						
MSCI classification	Developed	0.88***	0.86***	0.73***	0.84***	0.74***
		(9.29)	(8.84)	(6.49)	(8.03)	(6.27)
MSCI classification	Emerging	0.75***	0.71***	0.43	0.71***	0.46
		(3.04)	(2.94)	(1.51)	(2.71)	(1.52)
Panel B: Regulation Index						
Regulation Index	High	1.08***	1.05***	0.91***	0.87***	0.8***
		(8.51)	(7.94)	(5.75)	(5.93)	(5.02)
Regulation Index	Low	0.55***	0.5***	0.58***	0.57***	0.63***
		(3.2)	(2.84)	(2.5)	(2.8)	(2.49)
Panel C: Regulation proxies						
Criminal Sanctions	High	1.05***	1.03***	0.84***	0.84***	0.73***
		(7.17)	(6.63)	(4.87)	(5.05)	(4.09)
Criminal Sanctions	Low	0.58***	0.55***	0.59***	0.51***	0.55***
		(5.43)	(5.13)	(5.13)	(4.26)	(4.47)
Disclosure Requirements	High	0.93***	0.91***	0.83***	0.82***	0.78***
		(8.4)	(8.04)	(5.98)	(6.37)	(5.4)
Disclosure Requirements	Low	0.77***	0.69***	0.6***	0.75***	0.68***
		(4.21)	(4.28)	(3.25)	(4.62)	(3.68)
Investigative Powers	High	1.06***	1.04***	0.93***	0.91***	0.85***
		(8.9)	(8.18)	(6.05)	(6.64)	(5.61)
Investigative Powers	Low	0.67***	0.6***	0.56***	0.65***	0.59***
		(3.83)	(3.48)	(2.61)	(3.21)	(2.54)
Liability Standard	High	0.95***	0.93***	0.84***	0.87***	0.81***
		(8.58)	(7.94)	(6.13)	(6.6)	(5.58)
Liability Standard	Low	0.47***	0.37***	0.31**	0.23*	0.21
		(3.39)	(2.93)	(2.23)	(1.72)	(1.53)
Orders	High	1.01***	0.97***	0.82***	0.8***	0.72***
		(8.15)	(7.62)	(5.46)	(5.47)	(4.57)
Orders	Low	0.64***	0.57***	0.62***	0.65***	0.68***
		(3.74)	(3.22)	(2.64)	(3.19)	(2.6)
Public Enforcement	High	1.05***	1.02***	0.89***	0.87***	0.8***
		(8.56)	(8.0)	(5.82)	(6.17)	(5.18)
Public Enforcement	Low	0.49***	0.42***	0.46**	0.45**	0.47**
		(2.94)	(2.47)	(2.26)	(2.21)	(2.1)
Rule-Making Power	High	1.0***	0.96***	0.73***	0.86***	0.67***
		(6.65)	(6.01)	(4.5)	(5.36)	(4.05)
Rule-Making Power	Low	0.79***	0.73***	0.6***	0.7***	0.61***
		(5.47)	(5.3)	(3.75)	(3.77)	(3.26)
Supervisor Characteristic	High	1.27***	1.3***	1.06***	1.29***	1.08***
		(8.86)	(9.37)	(6.48)	(8.34)	(6.25)
Supervisor Characteristic	Low	0.57***	0.51***	0.39***	0.4**	0.34*
		(3.97)	(3.63)	(2.48)	(2.02)	(1.72)

We separately investigate developed and emerging markets as classified by MSCI as well as more and less regulated markets using proxies obtained from La Porta et al. (2006). We further create an regulation index using the average rank of a country based on these proxies. We report one-, three-, four-, five- and six-factor alphas based on *simBERT* driven long-short investments. The US is excluded from the analysis. The holding period is 12 months and we evaluate portfolios from January 1998 until June 2021. We denote the t-statistics of the coefficients in parentheses. * indicates significance at the 10% level, ** indicates significance at the 5% level and *** indicates significance at the 1% level.

Table 8: Analyst's reaction to annual report announcements

	World			US			W-US		
Dep Var	(1) [t-3;t+3]	(2) [t-12;t-3]	(3) [t+3;t+12]	(4) [t-3;t+3]	(5) [t-12;t-3]	(6) [t+3;t+12]	(7) [t-3;t+3]	(8) [t-12;t-3]	(9) [t+3;t+12]
EPS 1Y	10.63***	8.31***	0.58	36.53***	37.69***	5.37	5.36**	5.29	1.62
	(3.02)	(3.04)	(0.27)	(4.90)	(3.54)	(0.63)	(2.35)	(1.50)	(0.61)
EPS 2Y	8.94**	6.31**	1.21	28.96***	28.96***	9.24	4.36	3.83	1.95
	(2.43)	(2.05)	(0.41)	(5.69)	(3.93)	(1.07)	(1.60)	(1.15)	(0.70)
$\Delta Recom$	3.72***	-0.41	-1.62	3.15	1.80	4.88	3.03**	-1.42	-3.57
	(2.71)	(-0.22)	(-0.61)	(0.87)	(0.57)	(1.11)	(2.08)	(-0.66)	(-1.41)

We obtain analyst earnings forecasts from IBES and calculate the amount of positive and negative analyst revisions on the firm level within a given month. We then relate the difference between positive and negative revisions to the amount of total revisions within a specific time horizon to obtain an analyst revision score. We then regress this analyst revision score on *simBERT* and report the t-statistics as well as the coefficients in percentage points. More specifically, column (1), (4) and (7) report the coefficient obtained when considering revisions three months prior and post publication date. Columns (2), (5) and (8) report the coefficient for twelve months before up to three months prior to the publication date and columns (3), (6) and (9) for a period of three months until twelve months post publication date. Note that we consider global investments as well as investments into the US and and international markets outside the US separately. We control for year and firm fixed effects. We denote the t-statistics of the coefficients in parentheses. * indicates significance at the 10% level, ** indicates significance at the 5% level and *** indicates significance at the 1% level.

Table 9: **Median split on analyst coverage and market capitalization:**

Analyst	Size	1F	3F	4F	5F	6F
Panel A: Intl.						
Few		1.05***	0.96***	0.88***	0.95***	0.89***
		(7.11)	(6.86)	(4.83)	(5.56)	(4.59)
Many		0.61***	0.53***	0.46***	0.47***	0.43***
		(5.58)	(4.78)	(3.59)	(3.93)	(3.38)
Few	Small	0.98***	0.97***	0.9***	0.86***	0.83***
		(6.06)	(6.05)	(4.54)	(4.47)	(3.92)
Many	Small	0.96***	0.94***	0.86***	0.8***	0.77***
		(5.96)	(5.87)	(4.4)	(4.21)	(3.69)
Few	Large	0.38***	0.26***	0.22***	0.26***	0.23**
		(4.52)	(3.36)	(2.52)	(2.8)	(2.32)
Many	Large	0.39***	0.29***	0.24***	0.28***	0.25***
		(4.9)	(3.88)	(2.86)	(3.23)	(2.65)
Panel B: US						
Few		0.83***	0.85***	0.77***	0.72***	0.66***
		(5.72)	(6.56)	(5.57)	(5.8)	(5.03)
Many		0.35***	0.38***	0.3***	0.2*	0.13
		(2.67)	(3.38)	(2.62)	(1.76)	(1.15)
Few	Small	0.92***	0.9***	0.81***	0.66***	0.6***
		(4.4)	(4.59)	(4.06)	(3.12)	(2.82)
Many	Small	0.68***	0.71***	0.49**	0.5***	0.31
		(3.36)	(3.7)	(2.32)	(2.62)	(1.46)
Few	Large	0.4**	0.44***	0.38***	0.32**	0.28**
		(2.42)	(3.18)	(2.56)	(2.31)	(1.93)
Many	Large	0.34**	0.37***	0.29**	0.19	0.12
		(2.21)	(2.63)	(1.92)	(1.29)	(0.82)

We apply a median split based on analyst coverage and report one-, three-, four-, five- and six-factor alphas based on *simBERT* driven long-short investments for the different subgroups. We further control for market size by applying a double median split based on size and analyst coverage. The holding period is 12 months and we evaluate international and US portfolios from January 1998 until June 2021. We denote the t-statistics of the coefficients in parentheses. * indicates significance at the 10% level, ** indicates significance at the 5% level and *** indicates significance at the 1% level.