Textual Changes in 10-Ks and Stock Price Crash Risk: Evidence from Neural Network Embeddings

ARTICLE HISTORY

Compiled May 25, 2023

ABSTRACT

Previous research attributes stock price crash risk to managerial bad news hoarding. Contrary to this notion, we find evidence that stock price crash risk is determined by investor inattention to textual changes in corporate disclosures. Using a large sample of 10-K filings, we estimate neural network embeddings to quantify the degree of textual changes in successive 10-Ks. We find that changes in 10-Ks have a positive and economically meaningful impact on one-year-ahead stock price crash risks. Our results suggest that investor inattention to textual changes in 10-Ks can have broader capital market consequences than previously documented.

KEYWORDS

Stock Price Crash Risk; Textual Disclosures; 10-K Changes; Investor Inattention; Deep Learning

1. Introduction

Stock price crashes represent abrupt and large declines in stock prices that have severe effects on investor welfare. Burgeoning research attributes stock price crashes to managerial bad news hoarding (e.g., Hutton, Marcus, & Tehranian, 2009; Jin & Myers, 2006; J.-B. Kim, Li, & Zhang, 2011). These studies suggest that, due to managerial incentives, managers have the tendency to hide adverse firm information which leads to bad news being stockpiled within the firm. Once the accumulated bad news reaches a tipping point where it can no longer be contained, investors will abruptly correct prices, thus leading to a stock price crash. Contrary to this view, this study provides new evidence that stock price crash risk is also driven by investor inattention to textual changes in 10-Ks.

The Form 10-K required by the SEC is arguably one of the most comprehensive sources of information for investors to assess a company's market value. Larger year-over-year textual *changes* in corporate disclosures are associated with higher information content because larger changes imply a larger fraction of *new* information compared to previous disclosures (e.g., Hanley & Gerard, 2010). However, although the 10-K contains rich information, recent evidence suggests that investors barely read 10-K filings and that investors miss large and critical parts of its information (Cohen, Malloy, & Nguyen, 2020; Loughran & McDonald, 2017).

Given this evidence, how changes in 10-Ks are related to future crash risk is ex ante unclear. On the one hand, the additional information content provided by 'changers' may help investors to better assess firm value. Consequently, one would expect that larger changes in 10-Ks help mitigate stock price crash risk by providing new information to investors. On the other hand, if investors are largely inattentive to changes in 10-Ks and miss a large part of the new information, investors may fail to incorporate potentially value-relevant information into stock prices. In this case, firm values may deviate from fundamental levels and, thus, increase stock price crash risk.

To test these competing predictions, we collect a large sample of 10-K filings retrieved from the SEC Online EDGAR system. To measure changes in 10-Ks, we use *Doc2Vec*, a neural network proposed by Le and Mikolov (2014) to convert the 10-K filings into so-called document embeddings (i.e., vector representations of documents) that preserve the documents' semantic information. We train the model on a large database of 10-K filings and calculate cosine similarities of successive 10-K filings to estimate their textual changes over years. Using of sample of 31 195 firm-years over the period from 1994-2018, we find that year-over-year changes in 10-Ks are associated with higher one-year-ahead stock price crash risk, even after controlling for other known crash determinants. These findings support the view that investor inattention towards changes in 10-K filings increase future stock price crash risk.

In additional analyses, we examine whether 10-K changes could still be associated with managerial bad news hoarding. Therefore, we estimate the fraction of negative words in the Management Discussion and Analysis (MD&A) as an intuitive measurement of manager tendencies to block negative news flow. Because the MD&A is written by senior managers, word choices in this section are likely to reflect managerial manipulation attempts (Lo, Ramos, & Rogo, 2017). However, we find no association between changes in 10-Ks and negative word choices in the MD&A. Instead, our results suggest that these two constructs are distinct, indicating that the effects of textual changes in 10-Ks on future crash risk are likely to follow different economic mechanisms than bad news hoarding.

Finally, to further strengthen the view that our results are indeed driven by investor inattention, we examine market reactions around 10-K filing dates. If investors are inattentive to changes in 10-Ks, we would expect that such changes only lead to weak market reactions around the filing date. Consistent with the results documented in Cohen et al. (2020), we find no short-term announcement effects of 10-K changes, nor do 10-K changes appear to affect investors' postdisclosure risk perceptions. However, we find a significant lagged market reaction over the four months following the initial filing which is consistent with our baseline results and the view that investors initially miss a large part of the information in 10-Ks.

Finally, to assists investors in *how* they can best screen 10-Ks for important textual changes, we compare the neural network approach *Doc2Vec* with traditional Bag-Of-Words (BOW) approaches. Our results show that 10-K changes estimated based on *Doc2Vec* document embeddings are relatively more informative in predicting future crash risk compared to traditional BOW approaches. This finding complements recent research suggesting that machine learning approaches are superior in analyzing textual content (e.g., El-Haj, Rayson, Walker, Young, & Simaki, 2019; Frankel, Jennings, & Lee, 2022; Huang, Wang, & Yang, 2022).

The contribution of our study is threefold. First, we add to the literature examining the determinants of future stock price crash risks. While prior work has mainly focused on the role of quantitative disclosures in determining stock price crash risks (e.g., Hutton et al., 2009; J.-B. Kim & Zhang, 2016), only a few studies examine qualitative disclosures in this context. Specifically, prior work shows that linguistically more complex financial disclosures are associated with higher crash risk (Ertugrul, Lei, Qiu, & Wan, 2017; C. F. Kim, Wang, & Zhang, 2019). We add to this literature by examining textual changes in 10-Ks and find that textual changes are incrementally informative to common crash determinants about future stock price crash risk. Thereby, we provide evidence contrary to the notion that stock price crashes are determined by bad news hoarding (e.g., Hutton et al., 2009; Jin & Myers, 2006). Instead, our findings imply that investor inattention towards important company disclosures increase future crash risk. So far, this explanation has only drawn very little attention (Safdar, Neel, & Odusami, 2022).

Second, we contribute to the literature that examines the information role of corporate disclosures, and in particular 10-K filings. Prior literature has documented that while investors react to textual changes in 10-Ks, this effect has become weaker over time (Brown & Tucker, 2011; Feldman, Govindaraj, Livnat, & Segal, 2010). While these studies attribute their findings to decreasing informativeness of 10-Ks, we find that changes in 10-Ks are informative about future stock price crash risk. Our findings are in line with more recent research such as Cohen et al. (2020) who find that although 10-Ks provide rich information, investors may fail to adequately incorporate the information into contemporary stock prices.

Third, we contribute to the literature of textual analysis in accounting and finance. Recent studies emphasize that the accounting and finance literature is well behind the curve in terms of natural language processing sophistication and that methods based on machine learning largely outperform traditional dictionary-based approaches (e.g., El-Haj et al., 2019; Frankel et al., 2022; Huang et al., 2022). We add to this literature by documenting the superiority of neural network embeddings compared to traditional BOW methods when estimating the similarity of corporate disclosures. Specifically, we find that 10-K changes measured using *Doc2Vec* are significantly more informative about future stock price crashes compared to traditional BOW measures that are predominantly used in the literature.

2. Hypotheses Development

The Form 10-K required by the SEC is arguably one of the most comprehensive sources of firm information available to investors (Luo, Li, & Chen, 2018). While 10-Ks contain important accounting numbers, the textual content makes up approximately 80% of the entire report (Lo et al., 2017). Previous research shows that textual content is incrementally informative to quantitative information, as changes in textual content predict significant market reactions (Cohen et al., 2020; Feldman et al., 2010). While most prior work has focused on short-term market reactions to textual changes in 10-Ks, little is known about whether and how changes in 10-Ks are related to further reaching capital market consequences, such as stock price crash risk.

A large strand of the literature attributes increasing crash risks to managers' information hoarding (e.g., C. F. Kim et al., 2019; Kothari, Shu, & Wysocki, 2009). This opaque behavior (asymmetric information distribution) of managers in disclosure decisions leads to an increased risk of mispricing, a left-skewed return distribution, and thus an increase in crash risk (Hutton et al., 2009; Jin & Myers, 2006; McNichols, 1988). By withholding information for an extended period of time, a tipping point is reached at which it can no longer be withheld. Once this new information is revealed to the market, stock prices can crash (Jin & Myers, 2006). However, because changes in 10-Ks signal new information, investors may use this information to value stocks more accurately, thereby reducing crash risk. Therefore, one would expect that changes in annual reports would help reduce the risk of a stock crashing by providing new information to investors. Hence, our first hypothesis is as follows:

HYPOTHESIS 1: *Textual changes in 10-Ks decrease future stock price crash risk.*

On the other hand, Cohen et al. (2020) show that investors are inattentive to changes in 10-Ks and miss a large part of their information and Loughran and McDonald (2017) note that daily EDGAR requests for 10-Ks are surprisingly low, with only 28.4 times downloaded by investors from the SEC website. Moreover, R. Li, Wang, Yan, and Zhao (2019) find that when pre-announcement investor attention predicts less surprised stock market reactions and weaker post-earnings announcement drifts. Likewise, as a result of increased investor attention, Tao (2017) documents a significant decline in index returns. If 10-K changes signal information but investors do, on average, fail to incorporate this information, 10-K changes may lead to mispricing, thereby increasing stock price crash risk. Ni, Peng, Yin, and Zhang (2020) document a positive relationship between shareholder distraction and crash risk, linking inattention to crash risk. Thus, our second hypothesis is as follows:

HYPOTHESIS 2: *Textual changes in 10-Ks increase future stock price crash risk.*

3. Sample selection, data, and research design

3.1. Sample and data sources

After applying the screens and filters suggested by Schmidt, Schrimpf, von Arx, Wagner, and Ziegler (2019), our sample includes all available US firms between 1993 and 2019 from *Refinitiv Datastream* and *Refinitiv Worldscope.* Following previous research such as C. F. Kim et al. (2019), we exclude firms from highly regulated industries (SIC codes 4900–4999 and 6000–6999), remove firm-years with nonpositive total assets, negative market-to-book values, and require year-end stock prices to be greater than or equal to 1$. Moreover, we require at least 26 weekly returns to estimate our crash risk measures. To examine textual changes in 10-Ks, we retrieve all available 10-K filings from the Stage One Parsed 10-X database provided by Loughran and McDonald that are primarily sourced from the SEC's EDGAR system.[1] After requiring available data to construct control variables, we are left with 31 195 firm-years over the period from 1994-2018. However, the sample size for each regression model differs, depending on the data availability for all included variables.

3.2. Measuring 10-K similarity

Traditionally, the accounting and finance literature uses BOW models to estimate similarities between textual disclosures (e.g., Cohen et al., 2020; Hoberg & Maksimovic, 2014; Loughran & McDonald, 2016). The BOW model populates a vector with word counts of each unique word that is found in a set of documents. This vector serves as a numerical representation of a given document which can then be used to estimate document similarities by calculating distances between documents in the vector space. However, research in the field of computational linguistic points out several drawbacks of this method. First, BOW approaches do not account for word order. Different documents can convert into similar vectors as long as the same words are used. Second, BOW vectors do not account for semantic relations. For example, considering the words 'profit', 'sales', and 'dog', the BOW approach would consider all three words to be equally related to each other. However, the words 'profit' and 'sales' are clearly more closely related than the words 'profit' and 'dog'.

To overcome these obstacles, we use a machine learning approach to estimate year-over-year changes in 10-Ks. Specifically, we employ the *Doc2Vec* model developed by Le and Mikolov (2014), a neural network that aims to estimate vector representations of documents that capture the documents' semantic information. Therefore, the model builds on the *Word2Vec* model proposed by Mikolov, Corrado, Chen, and Dean (2013) that relies on an old rule in linguistics: words that co-occur with similar word-neighbors likely share a common meaning (Harris, 1954). In this sense, the neural network 'reads'

[1] The data is available from `https://sraf.nd.edu/data/stage-one-10-x-parse-data/`.

through a set of document and estimates vector representations of words based on their word-neighbors. As a result, words with similar meanings will occupy close locations in a vector space. *Doc2Vec* builds on this approach to estimate vector-representations of full documents rather than single words.[2]

We implement *Doc2Vec* using the *gensim* library in Python. We train the model based on all available 10-Ks from the Stage One Parsed 10-X database provided on the homepage of Loughran and McDonald. In summary, we collect 249 470 10-K filings. We preprocess the 10-K filings by (i) employing stemming to reduce feature dimensionality, (ii) forming phrases in the spirit of Mikolov, Sutskever, Chen, Corrado, and Dean (2013), (iii) and removing stopwords. After preprocessing, we use the full set of 10-Ks to train a *Doc2Vec* model.[3] After training, the model assigns each 10-K a 'learned' document embedding (i.e., a vector representation) that captures the document's semantic information. To measure the similarity of two successive 10-Ks, we estimate cosine similarities between two 10-Ks' document embeddings. The cosine similarity ranges between 0 and 1 with higher (lower) values indicating greater similarity (changes) between two 10-Ks.

3.3. Measures of stock price crash risk

To measure firm-specific stock price crash risk, we first estimate the following extended market model for each firm and fiscal year (e.g., C. F. Kim et al., 2019; J.-B. Kim et al., 2011; J.-B. Kim, Wang, & Zhang, 2016):

$$r_{i\tau} = \alpha_i + \beta_{1i} r_{m\tau-1} + \beta_{2i} r_{j\tau-1} + \beta_{3i} r_{m\tau} + \beta_{4i} r_{j\tau} + \beta_{5i} r_{m\tau+1} + \beta_{6i} r_{j\tau+1} + \epsilon_{i\tau} \qquad (1)$$

where r_i, r_m and r_j are the return in week τ for Stock i, the CRSP value-weighted market index m, and the Fama-Frech value-weighted index for industry j. Following C. F. Kim et al. (2019), we define a fiscal year as the 12 months ending three months after the fiscal year-end to avoid look-ahead bias, as 10-K reports are usually filed within three months after the fiscal year-end. We extend the index model with lead and lag terms for market, and industry returns to account for non-synchronous trading in our estimation (Dimson, 1979). To calculate the company-specific weekly returns for firm i in week τ, $W_{i\tau}$, we take natural logarithm of one plus the residual from equation 1.

Following the extant literature on stock price crash risks (Hutton et al., 2009; Je-

[2] For a detailed description see Le and Mikolov (2014).

[3] We estimate use the Distributed Bag Of Words (DBOW) method to train our model instead of the Distributed Memory (DM) method because the DBOW method is computationally less demanding. The hyperparameters are the number embedding dimensions (300), the considered word window (5), minimal number of words to be considered for training (50), negative sampling (5), and the number of epochs used for training (20). The remaining parameters are default parameters of the *gensim* library.

bran, Chen, & Zhang, 2021; Kao, Huang, Fung, & Liu, 2020; J. Kim, Li, & Zhang, 2011; J.-B. Kim et al., 2011; Y. Ma & Xu, 2021), we calculate the following three measures. First, we compute $NCSKEW_{it}$ which is calculated by taking the negative of the third moment of firm-specific weekly returns scaled by the standard deviation of firm-specific weekly returns raised to the third power. Formally, we calculate $NCSKEW_{it}$ as follows:

$$NCSKEW_{it} = \frac{-\left(n(n-1)^{\frac{3}{2}}\sum W_{i\tau}^{3}\right)}{\left((n-1)(n-2)(\sum W_{i\tau}^{2})^{\frac{3}{2}}\right)} \tag{2}$$

The second crash risk variable, $DUVOL_{it}$, is the asymmetric volatility of the weekly stock return. $DUVOL_{it}$ is calculated by taking the natural logarithm of the ratio of the standard deviation of firm-specific weekly returns in *down* weeks to the standard deviation of firm-specific weekly returns in *up* weeks, where the down (up) weeks are those with firm-specific weekly returns below (above) their mean value in fiscal year t. Formally, $DUVOL_{it}$ is calculated as follows:

$$DUVOL_{it} = \log\left\{\frac{(n_u-1)\sum_{DOWN} W_{i,\tau}^{2}}{(n_d-1)\sum_{UP} W_{i,\tau}^{2}}\right\}, \tag{3}$$

where n_u and n_d represent the number of up and down weeks over the fiscal year t, respectively. Higher values of $DUVOL_{it}$ correspond to higher crash risk.

Our third measure is $COUNT_{it}$ and is defined as the difference in frequencies between positive upward stock price jumps and negative stock price crashes in firm-specific returns. Stock price crashes (upward jumps) are firm-specific weekly return that falls (rises) 3.09 standard deviations below (above) the annual mean. We calculate $COUNT_{it}$ as the difference between stock price crashes and upward jumps (Jin & Myers, 2006). Higher values of $COUNT_{it}$ indicate increased stock price crash risk for firm i in year t.

3.4. Empirical Model

To test our hypotheses, we estimate the following model:

$$CRASH\_RISK_{i,t+1} = \beta_0 + \beta_1 SIMILARITY_{i,t} + \sum_k \beta_k CONTROLS_{i,t}^{k} + \epsilon_{i,t} \tag{4}$$

where $CRASH\_RISK_{it+1}$ is measured as one of the crash measures $NCSKEW_{it}$,

$DUVOL_{it}$ or $COUNT_{it}$ in fiscal year $t + 1$. The key independent variable, $SIMILARITY_{it}$ in fiscal year t is the cosine similarity between the 10-K filed in fiscal year $t - 1$ and the 10-K filed in fiscal year t. The similarities are estimated based on *Doc2Vec* document embeddings as described in section 3.2. Consistent with prior studies, we control for several crash risk determinants (Hutton et al., 2009; J.-B. Kim, Si, Xia, & Zhang, 2021). These include the three-year moving sum of absolute value of abnormal accruals ($OPAQUE_{it}$) and its squared term to account for a potentially non-linear relation between accruals management and crash risk. Further, we control for a set of firm fundamentals, including firm size ($LOGMV_{it}$), market-to-book ratio (MTB_{it}), financial leverage (LEV_{it}), and return on assets ($EARN_{it}$). Moreover, we include detrended stock trading volume ($DTURN_{it}$), the negative skewness of firm-specific returns ($NCSKEW_{it}$), weekly firm-specific return volatility (STD_RET_{it}), and the average weekly firm-specific return (RET_{it}) as market-based controls that could affect crash risk. Finally, all regression including firm- and year-fixed effect to control for firm- and year-wide variation in crash risk patterns.

4. Results

4.1. Descriptive statistics

Table 1 presents the descriptive statistics of the main variables used in this study. The average cosine similarity between two documents between two years equals 0.67. The mean values of the crash risk measures $NCSKEW_{it+1}$, $DUVOL_{it+1}$ and $COUNT_{it+1}$ are 0.110, 0.011, and 0.004 with a standard deviation of 0.911, 0.458, and 0.688, respectively. The average firm in our sample is a growth firm, as indicated by a mean MTB_{it} value of 4.484, and has a leverage ratio of 0.453, as shown by the mean value of LEV_{it}. Collectively, our descriptive statistics are comparable to those of C. F. Kim et al. (2019), who use a similar sample selection procedure.

Table 2 shows the Pearson correlations between our main variables. Larger firms and growth firms are associated with higher future crash risk as indicated by the positive correlations between the crash risk measures and the control variables $LOGMV_{it}$ and MTB_{it}. Similarly, the level of earnings management ($OPAQUE_{it}$) is positively and statistically significantly correlated with $NCSKEW_{it}$ and $DUVOL_{it}$. Finally, the correlation between $SIMILARITY_{it}$ and the crash risk variable $COUNT_{it}$ is consistent with our prediction, albeit statistically insignificant.

4.2. Main results

Table 3 presents the results of an OLS regression of crash risk on textual changes in 10-Ks and control variables. We find that the coefficient of $SIMILARITY_{it}$ is negative and statistically significant on the 1%-level for all three crash risk measures, $NCSKEW_{it+1}$, $DUVOL_{it+1}$ and $COUNT_{it+1}$. This finding supports our second hypothesis, H2, suggesting that textual changes in 10-Ks increase future stock price crash risk. These results may suggest that investors do not fully incorporate textual changes in 10-Ks into asset pricing, thereby increasing the risk of severe mispricings.

We find that the association between textual changes in 10-Ks and future crash risk is economically meaningful. Specifically, we find that a one standard deviation increase in $SIMILARITY_{it}$ is associated with 1.8% lower one-year-ahead crash risk as measured in standard deviations of the $NCSKEW_{it+1}$ ($= \frac{-0.185 \times 0.091}{0.911}$) distribution.[4] Hence, the economic effect of 10-K changes on crash risk is comparable to other crash determinants such as the effects of textual information obfuscation or corporate customer concentration (C. F. Kim et al., 2019; X. Ma, Wang, Wu, & Zhang, 2020).[5] Collectively, we consider the effect of $SIMILARITY_{it}$ to be economically significant.

[4] Note that the results are similar across all three crash risk measures. Specifically, a one standard deviation increase in $SIMILARITY_{it}$ is associated with 1.5% lower one-year-ahead crash risk as measured in standard deviations of both the $DUVOL_{it+1}$ and $COUNT_{it+1}$ distribution.

[5] Using the descriptive statistics published in C. F. Kim et al. (2019) and X. Ma et al. (2020), a one-standard-deviation increase in readability (corporate customer concentration) is associated with 1.5% (1.1%) higher future crash risk, measured using $NCSKEW_{it}$.

Hence, our findings could help investors secure their investment performance against stock price crashes.

Turning to the control variables, we find that our results are similar to previous research. Specifically, we find that large firms ($LOGMV_{it}$) and growth firms (MTB_{it}) are more likely to experience crashes (e.g., Chen, Hong, & Stein, 2001; Hutton et al., 2009; C. F. Kim et al., 2019). Furthermore, we find a positive association between ROA_{it} which is consistent with the results of (C. F. Kim et al., 2019).

4.3. Additional Analyses

4.3.1. Bad news hoarding

A large strand of literature suggests that stock prices crashes are driven by bad news hoarding (e.g., Hutton et al., 2009; Jin & Myers, 2006; J. Kim et al., 2011; J.-B. Kim et al., 2011). If managers use their discretion over financial disclosures to limit bad news flow, bad news can stockpile within a firm and result in a stock price crash once the bad news are released to the market. However, our results suggest that the relation between 10-K changes and stock price crashes is driven by investor inattention rather than opportunistic management behavior.

At this point, one could raise the question of whether changes in 10-Ks could facilitate bad news hoarding. To test this notion, we conduct a mediation analysis in the spirit of Baron and Kenny (1986) to assess the extent to which the effect of 10-K changes on crash risk could be explained by another mediator variable that reflects bad news hoarding behavior. Following Reichmann, Möller, and Hertel (2021), we adopt the simple intuition that managers who try to block the negative news flow will intentionally or unintentionally use less negative language when discussing their fiscal year. Therefore, we estimate the fraction of negative words in the MD&A section of the 10-K. The MD&A is written by senior managers, and thus, is likely to reflect attempts of opportunistic management behavior (Lo et al., 2017).

In Table 5, we first run an OLS regression of negative words in the MD&A ($NEGW_{it}$) on 10-K changes. However, our results suggest that 10-K changes have no effect on negative news flow. In step 2 of the mediation analysis, we reproduce our baseline results, showing that 10-K changes affect future crash risk as shown in columns (2)-(4). Finally, in columns (5)-(7) we add $NEGW_{it}$ to the regression. Consistent with the bad news hoarding theory, the results show that more (less) negative news flow significantly decreases (increases) future crash risk. However, and more importantly, the coefficient of $SIMILARITY_{it}$ remain virtually unchanged after including $NEGW_{it}$. This result suggests that the effects of bad news hoarding behavior and changes in 10-Ks on future crash risk are distinct from each other and that both are likely to follow different economic mechanisms.

4.3.2. 10-K changes and announcement effects

To further strengthen our inference that 10-K changes affect future crash risk through investor inattention, we next examine the relation between 10-K changes and announcement returns. If investors are indeed inattentive to changes in 10-Ks, we would expect a weak or insignificant announcement effect during the trading days surrounding the 10-K release. To test this notion, we estimate cumulative abnormal returns (CARs) in a seven-day window around the release date [-3;+3]. In addition, we also examine larger event windows ([+4;+63]; [+4;+126]) to identify lagged market reactions. Finally, following Kravet and Muslu (2013), we examine the effects of 10-K changes on investors' postdisclosure risk perception, measured as the change in two-month stock volatility before and after the filing date. All regressions control for $LOGMV_{it}$, MTB_{it}, PRE_RMSE_{it}, $EARNVOL_{it}$, $SIGMA_{it}$, and $LOSS_{it}$. PRE_RMSE_{it} is the Root Mean Squared Error of a pre-filing market model using trading days [-252; -6] relative to the 10-K filing to control for information uncertainty (Loughran & McDonald, 2014). $EARNVOL_{it}$ is a firm's earnings volatility, defined as the standard deviation of ROA_{it} over the five fiscal years from $t-4$ to t.[6] $LOSS_{it}$ is a dummy variable that equals 1 if the firm reports a loss and 0 otherwise. All other variables are defined in section 3.4.

The results are presented in Table 4. Consistent with the results of Cohen et al. (2020) and the notion that investors are inattentive to changes in 10-Ks, we find no short-term announcement effect to 10-K changes (column (1)), nor do textual changes in 10-K affect investors' postdisclosure risk perception (column (4)). Instead, the economic effects of textual changes in 10-Ks only materialize in the long-run as indicated by the statistically significant association between 10-K changes and $CAR_{[+4;+126]}$. This observation is consistent with the view that, while 10-Ks contain important information, investors are inattentive to this information, and thus, underreact to the information content of 10-Ks around the filing date. Collectively, these findings support our inference that textual changes in 10-Ks affect future crash risk through investor inattention.

4.3.3. Comparison of similarity measures

As Sunder (2010) points out, stock price crashes risk can not be mitigated by portfolio diversification, only by screening. Hence, if investor inattention drives future stock price crashes it is important to investors to identify methods that help them to screen 10-Ks for potentially important changes. While most previous literature employs simple word count methods to convert documents into numerical representations that can be used to estimate document similarities (e.g., Cohen et al., 2020; Hoberg & Maksimovic, 2014; Loughran & McDonald, 2011, 2016), a growing stream of literature criticizes the strong reliance on these basic textual analysis techniques (e.g., El-Haj et

[6] We require at least two observations to estimate $EARNVOL_{it}$.

al., 2019; Frankel et al., 2022; Huang et al., 2022). These studies suggest that machine learning, and in particular deep learning, has opened new opportunities to conduct deeper and more informative analyses of textual content. Drawing on this strand of literature, we aim to examine the relative informativeness of our neural network embedding approach to calculate 10-K changes compared to more traditional methods frequently used in accounting and finance research.

We test two traditional methods that are frequently used to estimate document similarities. First, for the measure SIM_COUNT_{it}, we convert each 10-K into vector representations, where each vector dimension captures the word count of a unique word found in our sample of 10-Ks. Second, SIM_TFIDF_{it} is a similar measure, but instead of capturing raw word counts, each word count is multiplied by the word's term frequency-inverse document frequency (tf-idf) weight. As pointed out by Loughran and McDonald (2011), tf-idf assign words that occur very frequently in a set of documents lower weights because such words are likely to be less important to these documents. Both measures are common BOW techniques that do neither account for word order nor semantics.

In Table 6, we benchmark both traditional measures SIM_COUNT_{it} and SIM_TFIDF_{it} against the deep learning approach used in our main analysis ($SIMILARITY_{it}$). While the results generally confirm our main findings, columns (7)-(9) show that, when including all three measures in one model, the effects of $SIMILARITY_{it}$ consumes the effects of both traditional measures SIM_COUNT_{it} and SIM_TFIDF_{it}, which suggests that the deep learning approach is more likely to capture important textual changes in 10-Ks that are informative about future crash risk. This finding is consistent with recent research suggesting that more sophisticated machine leaning approaches are more likely to capture meaningful information from textual disclosures (e.g., El-Haj et al., 2019; Frankel et al., 2022; Huang et al., 2022).

5. Robustness tests

5.1. Propensity score matching

In this section, we conduct a series of robustness tests to examine the sensitivity of our findings. First, our inference that textual changes in 10-Ks increase future stock price crash risk may raise concerns about functional form misspecification (FFM). If the determinants of stock price crash risk are correlated with 10-K changes, the observed effects could merely be a reflection of changes in the determinants of crash risk. To address this concern, we perform a propensity score matching (PSM) analysis. Following Shipman, Swanquist, and Whited (2016), we use a 1:1 nearest neighbor matching without replacement and a caliper of 0.01. Propensity scores are estimated using a probit regression that predicts an indicator variable equal to 1 for firms with above median values for $SIMILARITY_{it}$ and 0 otherwise. The regression includes the determinants of future crash risk as well as year and firm fixed effects. By matching the propensity scores of firms with higher and firms with lower similarity scores so that the absolute value of the difference between the propensity scores within the same industry year is minimized, we construct a sample of 'twin' firms that have the same propensity to file 10-Ks that are relatively similar to the previous year's report.

Panel A in Table 7 presents the test diagnostics of the matched sample. The results show that while high- and low 10-K similarity groups differ significantly in one-year-ahead crash risk, the determinants of crash risk do not differ significantly between these two groups. This finding suggests that our inferences are not merely a reflection of differences in crash determinants between high and low 'changers'. To further alleviate concerns about FFM, we reproduce our baseline results using the matched sample. The results in Panel B of Table 7 suggest that our findings remain robust.

5.2. Omitted variables

Second, we test the sensitivity of our results to adding additional control variables that previous research found to significantly influence crash risk. Following C. F. Kim et al. (2019), we add additional business risk measures including cash flow volatility ($CFVOL_{it}$), sales volatility ($SALESVOL_{it}$), and earnings volatility ($EARNVOL_{it}$) as well as the Herfindahl–Hirschman index (HHI_{it}) based on 3-digit SIC codes to control for business competition. Moreover, we control for several textual characteristics that relate to managerial obfuscation. This includes the modified FOG Index ($MODFOG_{it}$) proposed by C. F. Kim et al. (2019) as a proxy for textual obfuscation, the fraction of weak modal words (WMW_{it}) as a proxy for ambiguous language (Ertugrul et al., 2017), and the fraction of negative words $NEGW_{it}$ as a more direct measure of negative information flow. In addition, we include three additional measures for firm's inherent information asymmetry based on the findings of Wu and Lai (2020), namely, intangible intensity $ADJROTA_{it}$ as suggested by Clausen and Hirth

(2016), firm's proprietary costs $PROP\_COST_{it}$ measured as a firm's R&D expense scaled by total assets, and firm age AGE_{it} as a natural proxy for information asymmetry (F. Li, 2008). Finally, following Cohen et al. (2020), we estimate CEO and CFO changes from text in 10-Ks.[7] CEO_CHANGE_{it} and CFO_CHANGE_{it} are indicator variables that equal 1 if a 10-K contains information about a CEO or CFO change respectively, and 0 otherwise. As Table 8 shows, the coefficient of $SIMILARITY_{it}$ remains statistically significant, suggesting that omitted variable bias is no serious concern.

5.3. Falsification test

Finally, we conduct a falsification test in the spirit of Christensen, Hail, and Leuz (2016), Ljungqvist, Zhang, and Zuo (2017), and C. F. Kim et al. (2019). Specifically, in a first stage regression, we run OLS regressions of $NCSKEW_{it+1}$, $DUVOL_{it+1}$, and $COUNT_{it+1}$ on potential determinants of 10-K changes such as CEO_CHANGE_{it}, CFO_CHANGE_{it}, $NEGW_{it}$, $POSW_{it}$, $COMPW_{it}$, $MODFOG_{it}$ and obtain the predicted values of all three crash risk measures. CEO_CHANGE_{it} and CFO_CHANGE_{it} are indicator variables that equals 1 if CEO or CFO has changed and 0 otherwise. $NEGW_{it}$, $POSW_{it}$ and $COMPW_{it}$ are the fraction of negative, positive or complex words in 10-Ks. $MODFOG_{it}$ is the modified FOG Index proposed by C. F. Kim et al. (2019). In a second stage, we regress the predicted values of crash risk on our $SIMILARITY_{it}$ measure, while excluding the determinants from the first stage. Assuming that the observed or unobserved selection variables produce a false relationship between crash risk and similarity, the coefficients of $SIMILARITY_{it}$ should be similar to that in Table 3. However, the results in Table 9 show that the coefficients of $SIMILARITY_{it}$ are weak and statistically insignificant in the falsification test. These results indicate that omitted variable bias is not serious in our setting.

[7] Following Cohen et al. (2020), we parse the 10-K documents for mentions of CEO or CFO turnover. Specifically, we search for instances where a word from the set 'appoint', 'elect', 'hire', 'new', and 'search' and a word from the set 'CEO', 'CFO', 'Chief Executive Officer', and 'Chief Financial Officer' appear within the same sentence.

6. Conclusion

In this study, we examine whether and how changes in qualitative firm disclosures affect stock price crash risk. We find that textual changes in 10-Ks are positively associated with future crash risk. This finding supports the inattention hypothesis, suggesting that investors are inattentive to large changes in 10-Ks, thereby increasing stock price crash risk (Cohen et al., 2020). In line with this view, additional analyses suggest that the effect of 10-K changes on future crash risk follows different economic mechanisms than managerial bad news hoarding.

We contribute to the literature by showing that 10-K similarity has an impact on crash risk. Thereby, we provide a new and theoretically grounded perspective on the determinants of future crash risk. While most previous research suggests that managerial bad news hoarding drives crash risk, we provide new evidence that future crash risk is also determined by investor inattention to textual disclosures. Consistent with recent findings of Cohen et al. (2020), our results suggest that despite the valuable information content of 10-Ks, investors do not price this information into current stock prices. While some prior studies argue that financial reports have become less informative over time (Brown & Tucker, 2011; Feldman et al., 2010), we find that changes in 10-Ks are informative about future stock price crashes that have real welfare implications for investors. Moreover, we document that novel machine learning methods produce more informative measures compared to traditional BOW approaches when estimating textual changes in financial disclosures.

References

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182.

Brown, S. V., & Tucker, J. W. (2011). Large-Sample Evidence on Firms' Year-over-Year MD&A Modifications. *Journal of Accounting Research*, *49*(2), 309-346.

Chen, J., Hong, H., & Stein, J. C. (2001). Forecasting crashes: trading volume, past returns, and conditional skewness in stock prices. *Journal of Financial Economics*, *61*(3), 345-381.

Christensen, H. B., Hail, L., & Leuz, C. (2016). Capital-Market Effects of Securities Regulation: Prior Conditions, Implementation, and Enforcement. *The Review of Financial Studies*, *29*(11), 2885–2924.

Clausen, S., & Hirth, S. (2016). Measuring the value of intangibles. *Journal of Corporate Finance*, *40*, 110-127.

Cohen, L., Malloy, C., & Nguyen, Q. (2020). Lazy Prices. *The Journal of Finance*, *75*(3), 1371-1415.

Dimson, E. (1979). Risk measurement when shares are subject to infrequent trading. *Journal of Financial Economics*, *7*(2), 197-226.

El-Haj, M., Rayson, P., Walker, M., Young, S., & Simaki, V. (2019). In search of meaning: Lessons, resources and next steps for computational analysis of financial discourse. *Journal of Business Finance & Accounting*, *46*(3-4), 265-306.

Ertugrul, M., Lei, J., Qiu, J., & Wan, C. (2017). Annual Report Readability, Tone Ambiguity, and the Cost of Borrowing. *Journal of Financial and Quantitative Analysis*, *52*(2), 811-836.

Feldman, R., Govindaraj, S., Livnat, J., & Segal, B. (2010). Management's tone change, post earnings announcement drift and accruals. *Review of Accounting Studies*, *15*, 915-953.

Frankel, R., Jennings, J., & Lee, J. (2022). Disclosure sentiment: Machine learning vs. dictionary methods. *Management Science*, *68*(7), 5514-5532. Retrieved from `https://doi.org/10.1287/mnsc.2021.4156`

Hanley, K. W., & Gerard, H. (2010). The Information Content of IPO Prospectuses. *The Review of Financial Studies*, *23*(7), 2821–2864.

Harris, Z. S. (1954). Distributional Structure. *WORD*, *10*(2-3), 146-162.

Hoberg, G., & Maksimovic, V. (2014). Redefining Financial Constraints: A Text-Based Analysis. *The Review of Financial Studies*, *28*(5), 1312-1352.

Huang, A. H., Wang, H., & Yang, Y. (2022). Finbert: A large language model for extracting information from financial text†. *Contemporary Accounting Research*, *n/a*(n/a). Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1111/1911-3846.12832`

Hutton, A. P., Marcus, A., & Tehranian, H. (2009). Opaque financial reports, R2, and crash risk. *Journal of Financial Economics*, *94*(1), 67-86.

Jebran, K., Chen, S., & Zhang, R. (2021). Board social capital and stock price crash risk. *Review of Quantitative Finance and Accounting*.

Jin, L., & Myers, S. C. (2006). R2 around the world: New theory and new tests. *Journal of Financial Economics*, *79*(2), 257-292.

Kao, E. H., Huang, H.-C., Fung, H.-G., & Liu, X. (2020). Co-opted directors, gender diversity, and crash risk: evidence from China. *Review of Quantitative Finance and Accounting*, *55*, 461–500.

Kim, C. F., Wang, K., & Zhang, L. (2019). Readability of 10-K Reports and Stock Price Crash Risk. *Contemporary Accounting Research*, *36*(2), 1184-1216.

Kim, J., Li, Y., & Zhang, L. (2011). Corporate tax avoidance and stock price crash risk: Firm-level analysis. *Journal of Financial Economics*, *100*(3), 639-662.

Kim, J.-B., Li, Y., & Zhang, L. (2011). CFOs versus CEOs: Equity incentives and crashes. *Journal of Financial Economics*, *101*(3), 713-730.

Kim, J.-B., Si, Y., Xia, C., & Zhang, L. (2021). Corporate derivatives usage, information environment, and stock price crash risk*. *European Accounting Review*, *0*(0), 1-35.

Kim, J.-B., Wang, Z., & Zhang, L. (2016). CEO Overconfidence and Stock Price Crash Risk. *Contemporary Accounting Research*, *33*(4), 1720-1749.

Kim, J.-B., & Zhang, L. (2016). Accounting Conservatism and Stock Price Crash Risk: Firm-level Evidence. *Contemporary Accounting Research*, *33*(1), 412-441.

Kothari, S. P., Shu, S., & Wysocki, P. D. (2009). Do Managers Withhold Bad News? *Journal of Accounting Research*, *47*(1), 241-276.

Kravet, T., & Muslu, V. (2013). Textual risk disclosures and investors' risk perceptions. *Review of Accounting Studies*, *18*, 1088–1122.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st international conference on international conference on machine learning - volume 32* (p. II–1188–II–1196). JMLR.org.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, *45*(2-3), 221-247.

Li, R., Wang, X., Yan, Z., & Zhao, Z. (2019). Sophisticated Investor Attention and Market Reaction to Earnings Announcements: Evidence From the SEC's EDGAR Log Files. *Journal of Behavioral Finance*, *20*(4), 490-503.

Ljungqvist, A., Zhang, L., & Zuo, L. (2017). Sharing Risk with the Government: How Taxes Affect Corporate Risk Taking. *Journal of Accounting Research*, *55*(3), 669-707.

Lo, K., Ramos, F., & Rogo, R. (2017). Earnings management and annual report readability. *Journal of Accounting and Economics*, *63*(1), 1-25.

Loughran, T., & McDonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *Journal of Finance*, *66*(1), 35-65.

Loughran, T., & McDonald, B. (2014). Measuring Readability in Financial Disclosures. *Journal of Finance*, *69*(4), 1643-1671.

Loughran, T., & McDonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research*, *54*(4), 1187-1230.

Loughran, T., & McDonald, B. (2017). The Use of EDGAR Filings by Investors. *Journal of Behavioral Finance*, *18*(2), 231-248.

Luo, J.-h., Li, X., & Chen, H. (2018). Annual report readability and corporate agency costs. *China Journal of Accounting Research*, *11*(3), 187-212.

Ma, X., Wang, W., Wu, J., & Zhang, W. (2020). Corporate customer concentration and stock price crash risk. *Journal of Banking & Finance*, *119*(C), S0378426620301692.

Ma, Y., & Xu, L. (2021). Major government customers and stock price crash risk. *Journal of Accounting and Public Policy*, *40*(6), 106900.

McNichols, M. (1988). A comparison of the skewness of stock return distributions at earnings and non-earnings announcement dates. *Journal of Accounting and Economics*, *10*(3), 239-

273.

Mikolov, T., Corrado, G., Chen, K., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *Advances in Neural Information Processing Systems*, 1-12.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 3111–3119.

Ni, X., Peng, Q., Yin, S., & Zhang, T. (2020). Attention! Distracted institutional investors and stock price crash. *Journal of Corporate Finance*, *64*(C), S0929119920301450.

Reichmann, D., Möller, R., & Hertel, T. (2021). Nothing But Good Intentions – The Search for Equity and Stock Price Crash Risk. *Working Paper*.

Safdar, I., Neel, M., & Odusami, B. (2022). Accounting information and left-tail risk. *Review of Quantitative Finance and Accounting*, *58*, 1709-1740.

Schmidt, P. S., Schrimpf, A., von Arx, U., Wagner, A. F., & Ziegler, A. (2019). Common risk factors in international stock markets. *Financial Markets and Portfolio Management*(3), 213-241.

Shipman, J. E., Swanquist, Q. T., & Whited, R. L. (2016, 03). Propensity Score Matching in Accounting Research. *The Accounting Review*, *92*(1), 213-244.

Sunder, S. (2010). Riding the accounting train: From crisis to crisis in eighty years. Lehigh University, Bethlehem, PA: Presentation at the Conference on Financial Reporting, Auditing and Governance.

Tao, C. (2017). Investor Attention and Global Stock Returns. *Journal of Behavioral Finance*, *18*(3), 358-372.

Wu, K., & Lai, S. (2020). Intangible intensity and stock price crash risk. *Journal of Corporate Finance*, *64*(101682).

Table 1. Descriptive Statistics

Panel A: descriptive statistics

Variables	Observations	Mean	Std. Dev.	Q1	Median	Q3
Crash risk measures						
$NCSKEW_{it+1}$	31195	0.110	0.911	−0.419	0.031	0.528
$DUVOL_{it+1}$	31195	0.011	0.458	−0.278	−0.033	0.235
$COUNT_{it+1}$	31195	0.004	0.688	0.000	0.000	0.000
Independent Variables						
$OPAQUE_{it}$	31195	0.325	0.418	0.114	0.212	0.382
$LOGMV_{it}$	31195	13.324	2.160	11.797	13.344	14.811
MTB_{it}	31195	4.484	9.710	1.419	2.395	4.163
LEV_{it}	31195	0.453	0.223	0.279	0.457	0.612
ROA_{it}	31195	0.005	0.459	0.005	0.081	0.146
$DTURN_{it}$	31195	0.003	0.120	−0.020	0.002	0.029
$NCSKEW_{it}$	31195	0.093	0.900	−0.423	0.014	0.502
$SIGMA_{it}$	31195	0.055	0.035	0.032	0.046	0.069
RET_{it}	31195	−0.212	0.325	−0.235	−0.105	−0.049
$SIMILARITY_{it}$	31195	0.672	0.091	0.627	0.686	0.735

This table presents the descriptive statistics on crash risk, similarity, and controls. The crash variables cover the period 1994–2019, while the control variables cover the period 1993–2018. All variables are winsorized at the top and bottom 1%.

Table 2. Correlation Matrix

Variables		1	2	3	4	5	6	7	8	9	10	11	12	13
$NCSKEW_{it+1}$	1	1												
$DUVOL_{it+1}$	2	0.937***	1											
$COUNT_{it+1}$	3	0.774***	0.680***	1										
$OPAQUE_{it}$	4	0.033***	0.034***	0.000	1									
$LOGMV_{it}$	5	0.106***	0.130***	0.110***	−0.240***	1								
MTB_{it}	6	0.046***	0.048***	0.023***	0.171***	0.083***	1							
LEV_{it}	7	−0.002	−0.000	−0.001	−0.081***	0.244***	0.229***	1						
ROA_{it}	8	−0.014**	−0.015***	0.022***	−0.417***	0.261***	−0.241***	0.071***	1					
$DTURN_{it}$	9	0.010*	0.013**	0.008	−0.004	0.004	0.008	−0.002	0.006	1				
$NCSKEW_{it}$	10	0.182***	0.233***	0.089***	0.027***	0.074***	−0.004	0.003	−0.015***	0.001	1			
$SIGMA_{it}$	11	−0.156***	−0.231***	−0.113***	0.290***	−0.534***	0.062***	−0.110***	−0.298***	0.066***	−0.116***	1		
RET_{it}	12	0.096***	0.145***	0.091***	−0.292***	0.447***	−0.094***	0.068***	0.319***	−0.071***	0.093***	−0.928***	1	
$SIMILARITY_{it}$	13	0.008	0.019***	−0.003	−0.039***	−0.051***	−0.006	−0.031***	0.032***	−0.003	0.013**	−0.067***	0.069***	1

This table presents the Pearson correlations among the main variables in the study. All the crash risk and control variables are winsorized at the first and 99th percentiles. The 1%, 5%, and 10% significance levels of the coefficients are denoted by ***, **, and *, respectively.

Table 3. Similarity and Stock Price Crash Risk

	Dependent Variable		
	$NCSKEW_{it+1}$ (1)	$DUVOL_{it+1}$ (2)	$COUNT_{t+1}$ (3)
$SIMILARITY_{it}$	−0.185*** (−3.024)	−0.078*** (−2.824)	−0.114** (−2.311)
$OPAQUE_{it}$	0.008 (0.217)	0.002 (0.121)	0.028 (0.996)
$OPAQUE_{it}^2$	−0.002 (−0.155)	−0.003 (−0.562)	−0.002 (−0.258)
$LOGMV_{it}$	0.164*** (21.661)	0.082*** (24.017)	0.099*** (16.371)
MTB_{it}	0.001* (1.810)	0.001* (1.927)	0.001*** (2.639)
LEV_{it}	−0.048 (−1.174)	−0.030 (−1.625)	−0.034 (−1.053)
ROA_{it}	0.055*** (3.183)	0.029*** (3.889)	0.043*** (3.060)
$DTURN_{it}$	0.019 (0.504)	0.022 (1.248)	0.030 (0.970)
$NCSKEW_{it}$	−0.059*** (−9.045)	−0.026*** (−8.994)	−0.037*** (−7.287)
$SIGMA_{it}$	0.374 (0.610)	−0.024 (−0.088)	0.565 (1.158)
RET_{it}	0.042 (0.747)	0.022 (0.839)	0.087** (1.961)
Year fixed effects	Yes	Yes	Yes
Firm fixed effects	Yes	Yes	Yes
Observations	31 195	31 195	31 195
R^2	0.023	0.027	0.014

This table presents the results for the OLS regressions of crash variables $NCSKEW_{it+1}$, $DUVOL_{it+1}$ and $COUNT_{it+1}$ on our key independent variable $SIMILARITY_{it}$ for the time period from 1994 to 2018. The t-statistics reported in parentheses are based on White robust standard errors clustered by firm. All variables are winsorized at the first and 99th percentiles. The 1%, 5%, and 10% significance levels of the coefficients are denoted by ***, **, and *, respectively.

Table 4. Investor Response to 10-K Changes

	Dependent Variable			
	$CAR[-3,3]$	$CAR[4,63]$	$CAR[4,126]$	$\Delta\sigma(Return)$
	(1)	(2)	(3)	(4)
$SIMILARITYit$	−0.005	0.026	0.040**	−0.003
	−0.062	1.058	2.331	−0.037
$LOGMV_{it}$	−0.052***	−0.015***	−0.027***	−0.062***
	−4.626	−4.441	−10.951	−6.182
MTB_{it}	0.0002	0.001*	0.001***	0.001
	0.189	1.923	2.852	1.015
PRE_RMSE_{it}	−0.014	0.026***	0.020***	−0.211***
	−1.585	9.221	9.693	−19.174
$EARNVOL_{it}$	0.0005	0.001	−0.001	0.005
	0.089	0.757	−0.622	1.288
$SIGMA_{it}$	1.469***	0.249*	0.074	−2.290***
	3.168	1.790	0.756	−4.587
$LOSS_{it}$	−0.141***	−0.006	−0.009	0.111***
	−5.025	−0.746	−1.519	4.627
Year fixed effects	Yes	Yes	Yes	Yes
Firm fixed effects	Yes	Yes	Yes	Yes
Observations	31 908	31 908	31 908	31 908
R^2	0.002	0.009	0.017	0.051

This table presents the results for the OLS regression of Cumulative Average Returns (CAR) on our key independent variable $SIMILARITY_{it}$. All the dependent variables and controls are winsorized at the first and 99th percentiles. The 1%, 5%, and 10% significance levels of the coefficients are denoted by ***, **, and *, respectively.

Table 5. Bad News Hoarding – Mediating Effect of Negative Wording in the MD&A Section

	Step 1	Step 2			Step 3		
	$NEGW_{it}$ (1)	$NCSKEW_{it+1}$ (2)	$DUVOL_{it+1}$ (3)	$COUNT_{it+1}$ (4)	$NCSKEW_{it+1}$ (5)	$DUVOL_{it+1}$ (6)	$COUNT_{it+1}$ (7)
$SIMILARITY_{it}$	−0.000 (−0.743)	−0.192*** (−3.059)	−0.083*** (−2.892)	−0.122** (−2.388)	−0.193*** (−3.077)	−0.083*** (−2.913)	−0.122** (−2.403)
$NEGW_{it}$					−3.864*** (−3.130)	−2.157*** (−3.805)	−2.676*** (−2.589)
Other Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	30 005	30 005	30 005	30 005	30 005	30 005	30 005
R^2	0.065	0.025	0.030	0.015	0.026	0.030	0.015

This table presents the results for the mediation analysis as proposed by Baron and Kenny (1986). In “Step 1”, we regress the mediator $NEGW_{it}$ which is defined as the fraction of negative words in a given MD&A section, on our main independent variable $SIMILARITY_{it}$. In “Step 2”, we run OLS regressions of $NCSKEW_{it+1}$, $DUVOL_{it+1}$ and $COUNT_{it+1}$ on our similarity measure $SIMILARITY_{it}$. In “Step 3”, we also add the mediator $NEGW_{it}$ to the regression model. The t-statistics reported in parentheses are based on White robust standard errors clustered by firm. All variables are winsorized at the first and 99th percentiles. The 1%, 5%, and 10% significance levels of the coefficients are denoted by ***, **, and *, respectively.

Table 6. Alternative Measurement of Similarity

	Dependent Variable								
	$NCSKEW_{it+1}$	$DUVOL_{it+1}$	$COUNT_{it+1}$	$NCSKEW_{it+1}$	$DUVOL_{it+1}$	$COUNT_{it+1}$	$NCSKEW_{it+1}$	$DUVOL_{it+1}$	$COUNT_{it+1}$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
SIM_COUNT_{it}	−0.087*	−0.023	−0.038				0.095	0.059	0.113
	(1.858)	(−1.100)	(−1.010)				(0.892)	(1.212)	(1.304)
SIM_TFIDF_{it}				−0.091**	−0.030*	−0.052*	−0.132	−0.059	−0.116
				(−2.392)	(−1.714)	(−1.678)	(−1.533)	(1.519)	(−1.634)
$SIMILARITY_{it}$							−0.160**	−0.073**	−0.104**
							(−2.495)	(−2.514)	(−1.995)
Other Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	31 195	31 195	31 195	31 195	31 195	31 195	31 195	31 195	31 195
R^2	0.023	0.027	0.014	0.023	0.027	0.014	0.023	0.027	0.014

This table presents the results for the OLS regressions of $NCSKEW_{it+1}$, $DUVOL_{it+1}$ and $COUNT_{it+1}$ on different similarity measures SIM_COUNT_{it}, SIM_TFIDF_{it} and $SIMILARITY_{it}$. The t-statistics reported in parentheses are based on White robust standard errors clustered by firm. All variables are winsorized at the first and 99th percentiles. The 1%, 5%, and 10% significance levels of the coefficients are denoted by ***, **, and *, respectively.

Table 7. Robustness Test: Propensity Score Matching

Panel A: PSM results

	Treatment Variable: $SIMILARITY_{it}$			
Dependent Variables	High	Low	High−Low	t-stat.
$NCSKEW_{it+1}$	0.135	0.104	0.031***	2.757
$DUVOL_{it+1}$	0.024	0.007	0.017***	3.024
$COUNT_{it+1}$	0.017	−0.001	0.018**	2.112
Determinants of Crash Risk				
$OPAQUE_it$	0.311	0.308	0.003	0.729
$LOGMV_it$	13.308	13.323	−0.015	−0.591
MTB_it	4.291	4.252	0.039	0.411
LEV_it	0.455	0.453	0.002	0.663
ROA_it	0.021	0.024	−0.003	−0.991
$DTURN_it$	0.003	0.003	0.000	0.617
$NCSKEW_it$	0.097	0.095	0.002	0.234
$SIGMA_it$	0.055	0.055	0.000	0.080
RET_it	−0.204	−0.204	0.000	−0.121

Panel B: PSM firm fixed effects regression

	Dependent Variable		
	$NCSKEW_{it+1}$	$DUVOL_{it+1}$	$COUNT_{it+1}$
$SIMILARITY_{it}$	−0.225*** (−3.407)	−0.085*** (−2.831)	−0.163*** (−3.035)
Controls	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes
Firm fixed effects	Yes	Yes	Yes
Observations	27 356	27 356	27 356
R^2	0.024	0.028	0.015

This table presents the results for Propensity Score Matching (PSM) analysis. Panel A presents the test diagnostics. Panel B presents the results for the regression of crash measures $NCSKEW_{it+1}$, $DUVOL_{it+1}$, and $COUNT_{it+1}$ on $SIMILARITY_{it}$ using the PSM matched sample. The t-statistics reported in parentheses are based on White robust standard errors clustered by firm. All crash risk and control variables are winsorized at the first and 99th percentiles. The 1%, 5%, and 10% significance levels of the coefficients are denoted by ***, **, and *, respectively.

Table 8. Robustness Test: Additional Controls

	Dependent Variable		
	$NCSKEW_{it+1}$	$DUVOL_{it+1}$	$COUNT_{it+1}$
$SIMILARITY_{it}$	-0.186^{***}	-0.077^{**}	-0.111^{**}
	(-2.758)	(-2.530)	(-2.020)
$CFVOL_{it}$	0.044	0.020	0.015
	(1.540)	(1.404)	(0.641)
$SALESVOL_{it}$	-0.019	-0.011^{*}	-0.015
	(-1.471)	(-1.874)	(-1.461)
$EARNVOL_{it}$	-0.020	-0.009	-0.012
	(-1.432)	(-1.277)	(-1.099)
HHI_{it}	-0.135	-0.055	-0.065
	(-1.515)	(-1.370)	(-0.888)
$MODFOG_{it}$	0.009	0.003	0.003
	(1.537)	(1.042)	(0.678)
$WMOD_{it}$	-1.379	-0.814	-1.748
	(-0.475)	(-0.622)	(-0.736)
$ADJROTA_{it}$	0.019^{***}	0.010^{***}	0.019^{***}
	(3.099)	(3.493)	(3.950)
$PROP\_COST_{it}$	55.579	29.242	16.634
	(0.957)	(1.026)	(0.246)
AGE_{it}	-0.135^{**}	-0.027	-0.145^{***}
	(-2.165)	(-0.929)	(-2.798)
CEO_CHANGE_{it}	-0.025^{*}	-0.012^{*}	-0.019
	(-1.759)	(-1.824)	(-1.597)
CFO_CHANGE_{it}	0.011	0.006	0.029^{**}
	(0.659)	(0.839)	(2.149)
$NEGW_{it}$	-3.908^{**}	-2.023^{**}	-4.533^{***}
	(-2.093)	(-2.376)	(-2.977)
Other Controls	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes
Firm fixed effects	Yes	Yes	Yes
Observations	26 425	26 425	26 425
R^2	0.024	0.029	0.015

This table presents the results for the OLS regressions of $NCSKEW_{it+1}$, $DUVOL_{it+1}$ and $COUNT_{it+1}$ on our similarity measure $SIMILARITY_{it}$. Other controls are the same dependent variables as in Table 3. The t-statistics reported in parentheses are based on White robust standard errors clustered by firm. All variables are winsorized at the first and 99th percentiles. The 1%, 5%, and 10% significance levels of the coefficients are denoted by ***, **, and *, respectively.

Table 9. Robustness Test: Falsification Test

	Dependent Variable		
	$Pred\_NCSKEW_{it+1}$ (1)	$Pred\_DUVOL_{it+1}$ (2)	$Pred\_COUNT_{it+1}$ (3)
$SIMILARITY_{it}$	−0.001 (−0.102)	0.006 (1.535)	0.003 (1.156)
$OPAQUE_{it}$	0.003 (1.251)	0.002 (1.249)	0.001 (1.000)
$OPAQUE^2_{it}$	−0.001 (−1.614)	−0.001 (−1.536)	−0.0001 (−0.408)
$LOGMV_{it}$	0.003*** (4.816)	0.001*** (4.008)	0.003*** (12.426)
MTB_{it}	−0.000*** (−8.018)	−0.000*** (−7.629)	−0.000*** (−9.558)
LEV_{it}	0.037*** (13.692)	0.021*** (12.866)	0.022*** (18.361)
ROA_{it}	−0.002 (−1.277)	−0.001 (−1.482)	−0.000 (−0.349)
$DTURN_{it}$	−0.007*** (−2.779)	−0.004*** (−2.840)	−0.003** (−2.428)
$NCSKEW_{it}$	−0.001** (−2.011)	−0.000** (−2.094)	−0.000 (−1.073)
$SIGMA_{it}$	0.276*** (7.193)	0.162*** (7.056)	0.048*** (2.879)
RET_{it}	0.011*** (3.188)	0.006*** (2.923)	0.001 (0.468)
Year fixed effects	No	No	No
Firm fixed effects	No	No	No
Observations	31195	31195	31195
R^2	0.013	0.012	0.018

This table presents the results for the regressions of predicted $NCSKEW_{t+1}$, $DUVOL_{t+1}$, and $COUNT_t + 1$ on our key independent variable $SIMILARITY_{it}$. The t-statistics reported in parentheses are based on White robust standard errors clustered by firm. All variables are winsorized at the first and 99th percentiles. The 1%, 5%, and 10% significance levels of the coefficients are denoted by ***, **, and *, respectively.