Data Screens and Parsing

We divided the 10-Ks into their sections (Item 1–15) with regular expressions matching the headings. We removed exhibits, graphics, HTML tags, headings, and other anomalies leaving plain paragraph text. Lastly, following Loughran and McDonald (2011), we only kept sections with at least 250 words.

Similar to Loughran and McDonald (2014), we performed a set of data screens: After removing duplicates (dropping 3,301), we required a filing date of more than 180 days after the prior filing (dropping 653), a match with CRSP's permanent identifier PERMNO (dropping 113,396), the stock to be ordinary common equity (dropping 4,466), a stock price of greater than \$3 (dropping 15,281), a positive book-to-market (dropping 3,384), as well as stock return data available for trading day windows $[-252, -6]$ before, $[0, 1]$ during, and $[6, 28]$ after the filing date (dropping 409). Lastly, we removed reports in which we could not identify at least one complete section (dropping 2,684). This left us with 76,991 reports for financial regression analyses.

We took the residual of 126,330 files (excluding 17,244 duplicates and documents with less than one section) dropped during the screens and split it in two: 1,500 documents were randomly sampled for developing a classifier and the remainder of 124,830 files was used to train the embedding model.

All texts were tokenized, stripped of punctuation and numbers, and lowercased with the exception of proper nouns, which were identified through part-of-speech tagging.[1] Hence, we ensure that e.g. the modal verb "may" can be distinguished from the month "May".

[1] We used NLTK 3.2.1 for all of these steps.