Modeling Institutional Credit Risk with Financial News

Tam Tran-The

MassMutual Data Science
470 Atlantic Ave
Boston, Massachusetts 02210
ttranthe32@massmutual.com

Abstract

Credit risk management, the practice of mitigating losses by understanding the adequacy of a borrower's capital and loan loss reserves, has long been imperative to any financial institution's long-term sustainability and growth. MassMutual is no exception. The company is keen on effectively monitoring downgrade risk, or the risk associated with the event when credit rating of a company deteriorates. Current work in downgrade risk modeling depends on multiple variations of quantitative measures provided by third-party rating agencies and risk management consultancy companies. As these structured numerical data become increasingly commoditized among institutional investors, there has been a wide push into using alternative sources of data, such as financial news, earnings call transcripts, or social media content, to possibly gain a competitive edge in the industry. The volume of qualitative information or unstructured text data has exploded in the past decades and is now available for due diligence to supplement quantitative measures of credit risk. This paper proposes a predictive downgrade model using solely news data represented by neural network embeddings. The model standalone achieves an Area Under the Receiver Operating Characteristic Curve (AUC) of more than 80%. The output probability from this news model, as an additional feature, improves the performance of our benchmark model using only quantitative measures by more than 5% in terms of both AUC and recall rate. A qualitative evaluation also indicates that news articles related to our predicted downgrade events are specially relevant and high-quality in our business context.

1 Introduction

Credit risk refers to the possibility of loss resulting from a borrower's or a bond issuer's failure to repay a loan or meet contractual obligations. One subcategory of credit risk is downgrade risk, which occurs when third-party rating agencies, such as Moody's and S&P, lower their ratings on a bond or a company. For example, a change by S&P from a B to a CCC rating is considered a downgrade event. Such rating information is used extensively by regulatory organization, specifically the National Association of Insurance Commissioners (NAIC), to ensure financial solvency of insurance companies. More precisely, NAIC requires a company with a higher exposure to risk to hold a higher amount of capital in reserve. Accurately monitoring rating classes and potential deterioration events is therefore critical to MassMutual. Receiving enough notice before any possible impending downgrade event would help the company to manage its investment portfolio more efficiently by preparing necessary capital and/or switching holding positions.

In this paper, we design 3 natural language processing (NLP) frameworks to recognize credit relevant patterns (i.e. future 1-year downgrade events) in news articles about more than 2.2K companies of interest. The study demonstrates that with an appropriate methodology, we can achieve an AUC of more than 80% using solely news data to predict downgrade events and yield a more than 5% performance gain when this adverse credit signal in media coverage is added to the quantitative downgrade risk model.

The structure of this paper is as follows: Section 2 provides background on credit risk modeling and the NLP techniques used to convert text into meaningful numerical data; Section 3 describes the dataset, benchmark model and main model development; Section 4 presents performance results and stability test details of the final model; and Section 5 discusses limitations of this work and future directions.

2 Background

2.1 Credit Risk Modeling

The motivation to develop credit risk models stems from the need to construct quantitative estimates of the amount of economic capital needed to support a financial institution's risk-taking activities. Specifically, minimum capital in reserve is often set in proportion to the risk exposure of a company's portfolio. Typical credit risk models take as input the conditions of the general economy and those of the specific firms in question, and generate as output a credit quality measure.

2.2 Natural Language Processing (NLP) Techniques: Topic Modeling, Sentiment Lexicons and Neural Net Embeddings

When it comes to using text in a machine learning model, one of the main challenges is how to represent texts as nu-

merical inputs so that we can feed them into the model. This project aims to experiment with different methodologies to represent news articles and evaluate the algorithms' performances based on the downgrade prediction task. Following are some NLP techniques we use in the study:

Latent Dirichlet Allocation (LDA). As it is believed that every article we have is composed of major themes, we use topic modeling to extract the hidden thematic structure in the text. Topic modeling is a type of statistical model used to discover the abstract topics that occur in a collection of documents. Latent Dirichlet allocation (LDA), a generative probabilistic model, is a common algorithm of topic model. The basic idea is documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. Mathematically, this can be formulated as:

$$P(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta) = P(\theta | \alpha) \prod_{n=1}^{N} P(z_n | \theta) P(w_n | z_n, \beta)$$

Given α, a parameter vector on the per-document topic distributions and β, a parameter vector on the per-topic word distributions, we are finding the joint distribution of a topic mixture θ, a set of N topics $\mathbf{z}$, and a set of N words $\mathbf{w}$ (Blei et al. 2003). In the context of our study, each news article is represented as a set of topic probabilities and the number of topics is a tunable parameter. To evaluate an LDA model, we use coherence measure, which gives an estimate of how well each topic can be represented as a composition of parts that can be combined (Röder, Both, and Hinneburg 2015).

Sentiment Lexicons. There is a growing body of sentiment lexicons, or affective word lists, to examine the tone and sentiment of textual data. Some lexicons we experiment with are:

- **Loughran and McDonald**, designed to particularly reflect tone in financial texts (Loughran and McDonald 2011)

- **VADER**, specifically attuned to sentiment in microblog-like contexts (Hutto and Gilbert 2014)

- **AFINN**, also constructed based on micro-blog posts including Internet slang words (Nielsen 2011)

- **SentiWordNet** and **OPINION**, for general sentiment analysis

Neural Net Embeddings. Based on the underlying idea that "a word is characterized by the company it keeps," each word is represented by an embedding, or a vector of continuous numbers so that words of similar semantic (relative to the task) are closer to one another in the vector space. By using neural network on a supervised/unsupervised task, the resulting weights/parameters that have been adjusted to minimize loss on the task are the embeddings.

- **Doc2Vec Document Embeddings**. Doc2Vec is an unsupervised framework that learns fixed-length feature representations from variable-length pieces of texts. In the framework, a vector representation is trained by stochastic gradient descent and backpropagation to be useful for predicting the next word in the sentence. The Doc2Vec model we use in this study is a distributed memory one, where each paragraph vector acts as a memory that remembers what is missing from the current context (Le and Mikolov 2014). After being trained, the paragraph vectors can be used as features for the paragraph and be fed directly into conventional machine learning algorithms such as logistic regression.

- **fastText Word Embeddings**. Inspired by the same hypothesis as Doc2Vec, fastText word representations are trained to predict words that appear in their contexts. What distinguishes fastText model is it takes into account morphology. More precisely, the model represents each word by a sum of its character n-grams, which allows us to compute representations for words that did not appear in the training data and proves to be helpful while working with morphologically rich language (Bojanowski et al. 2016).

3 Modeling Downgrade Risk

3.1 Data

News. For each company at a point in time (daily), we have all news articles corresponding to the record and a label indicating whether the company is going to downgrade within future 1-year period.

- **Data source**. We have access to news data source from Thomson Reuters via CreditEdge's API provided by Moody's. This source brings us daily article pieces at company level. Reuters articles in the dataset could take various forms: news stories about company's activities (e.g. merger & acquisition, board assignment, etc.), market snapshots (e.g. individual stock movements, etc.), or quarterly earnings summaries. There are approximately 21K observations identified at company and date level, covering more than 2.2K Moody's unique permanent company identifiers (PID) and spanning from mid November 2017 to early October 2019 (although 99% of these articles are in 2019).

- **Data preprocessing**. Since we are working in a very narrow domain (financial news of a company of interest) and the data is particularly noisy, we take multiple cleaning and preprocessing steps. First, we remove small machine-generated articles, such as NYSE/NASDAQ/AMEX order imbalance, articles that only include video links or articles whose accuracy hasn't been verified by Reuters. Second, we remove unnecessary headers, footers, HTML metadata, and lowercase all texts. Third, we manually read through a handful of articles and flag which format tends to mention multiple companies in the text and which tends to talk about a single company. Last but not least, for pieces that cover information of multiple companies, we use a combination of named entity recognition and fuzzy matching techniques, as well as rule-based exact matching to extract only sentences about the company of relevance. A few more preprocessing layers, such as tokenization, stemming and stop words removal, are added depending on which modeling approach we take. By performing a thorough text preprocessing, we hope to
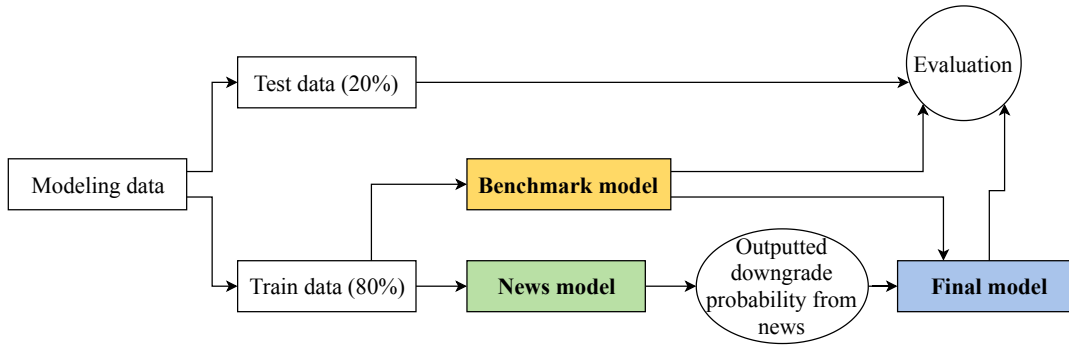
Figure 1: Overview of modeling pipeline. For details of benchmark model, see Section 3.2. For details of news model, see Section 3.3. Note that features for the final model include the outputted probability from news model and all quantitative features from benchmark model.

increase the signal to noise ratio which suggests news is more likely to have a material impact on our downgrade prediction. To ensure there is only one row of data per company per day, for companies that have multiple articles on a single day, we concatenate these articles together.

Ratings. Credit ratings for each company are available from both Moody's and S&P. We take the worst combination between the two rating systems for a company at a point in time (daily) to derive current rating and the worst rating within the next future 1 year, from which we determine whether that company has an impending downgrade event or not.

Table 1: Downgrade events by year

downgrade	year	count
	2017	42
0	2018	152
	2019	20,519
1	2019	324

For our benchmark model, all features are variations of quantitative credit measures provided by third-party rating agencies and risk management consultancy companies - this is further discussed in Section 3.2.

3.2 Benchmark Model

The benchmark downgrade model of this study is a logistic regression. For each training data point, we have a vector of features X and an observed class Y, where X provides quantitative metrics about a company on a daily basis and Y = 1 indicates that the company is downgrading within 1 year from the date of the data observation. This setup ensures that both our benchmark and news models have the same level of data frequency and can be compared and combined later. Assuming that $P(Y = 1|X = x) = p(x; \theta)$ for some function p parameterized by θ, what we want to model is:

$$P(Y = 1|X = x, \theta) = \frac{1}{1 + e^{-\theta x}}$$

The model is trained on 9 features that are variations in terms of either term structures (i.e. 1-year, 5-year) or transformation (i.e. lag, diff) of following quantitative measures purchased from risk management providers:

- The probability that a firm will default over 1 year based on company-specific attributes, industry-related measures and relevant macro-economic factors
- The probability that a firm will downgrade within 1 year based on market implied ratings and rating outlooks
- Historical credit rating for any firm at a point in time

Since our data is highly imbalanced, where the occurrence of downgrade events accounts for 1.5% of the dataset, we employ SMOTE algorithm to replicate observations from the minority class. Overall, this benchmark model achieves an AUC of 82.7% and a recall rate of 69.5%.

3.3 Model Development

Figure 1 is an overview of our modeling pipeline. Throughout the training process, we deliberately employ logistic regression due to the algorithm's great interpretability which is a focus of our business users who work in a highly regulated industry. This section details 3 different approaches we take to translate unstructured texts into meaningful numerical data that can be fed into a logistic regression model. We evaluate these NLP methodologies on our downstream task at hand, which is downgrade prediction, to select the best news model that can be incorporated into and improve the existing benchmark model.

Approach 1: Lexicon-based Sentiment and Topic Scores. We first experiment with one of the simplest sentiment analysis approaches, which is to compare words of an article against a labeled word list, where each word has been scored for valence. Based on 5 different lexicons (Loughran & McDonald, VADER, AFINN, SentiWordNet, and OPIN-ION), we compute 5 sentiment scores for each news article. Specifically, we check whether a word in the text exists in each positive or negative word list, then count the frequency of that word. For positive words, we do not count those that have a negation in one of three places preceding it. The sentiment score is calculated as $\frac{\text{#positives–#negatives}}{\text{#positives+#negatives}}$. Additionally,

we train an LDA model with 10 different topics where each topic is a combination of keywords and each keyword contributes a certain weight to the topic. We decide on 10 as the optimal number of topics by running multiple LDA models with number of topics ranging from 5 to 40 and picking the one that gives the highest coherence value, as shown in Figure 2. Two features that indicate the number of articles for each company at a point in time and whether an article contains any variation of the word "downgrade" are also created. In the end, our news model is trained on 17 features, including 5 sentiment scores, 10 topic probabilities, and 2 count indicators. This news model standalone has an unimpressive AUC of 59.8%. Although sentiment and topic scores are helpful in giving us a basic analysis of our news data at hand, these features are far from sufficient for accurate downgrade prediction.
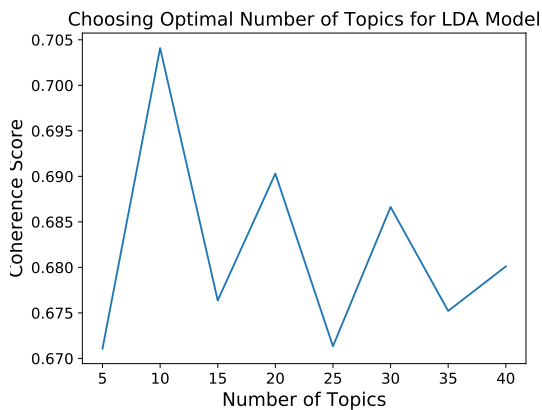

Figure 2: Coherence values of LDA model with different topic numbers. The optimal number is 10 topics with coherence value of 0.704.

Approach 2: Doc2Vec Embeddings. In this approach, we train a distributed-memory Doc2Vec model based on the vocabulary of our training dataset over 20 epochs and generate a document embedding for each news article. We choose to use Doc2Vec paragraph vectors since they are learned from unlabeled data and theoretically can work well for tasks that have small amount of labeled data (Le and Mikolov 2014), which is our case in this study. These embeddings are taken as input into a logistic regression model, which results in an AUC of 71.6%.

Approach 3: fastText Embeddings. Taking advantage of transfer learning, we use 1M word vectors pre-trained by a team of Facebook researchers on Wikipedia 2017, UMBC webbase corpus and statmt.org news dataset (Mikolov et al. 2018). Since the fastText model was trained on massive data sources, their representations perform very well at transferring to other NLP problems and improve the generalization of models learned on limited amount of data. Unlike Doc2Vec which produces representations at document level, fastText generates a 300-dimensional vector for each word. To create an embedding for the whole document, we take the average of embeddings for words contained in it. With
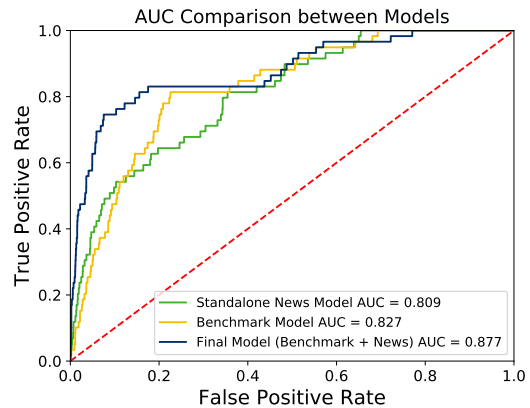

Figure 3: AUC comparison between benchmark, standalone news, and final models. Final model achieves the highest value, 87.7%, resulting in a 5% gain compared with benchmark model.

these final document embeddings fed into a logistic regression, our news model achieves an AUC of 80.9%, which is the best performance for this standalone model so far.

Table 2: Performance of standalone news model

approach	AUC (%)
sentiment scores and LDA topic scores	59.8
Doc2Vec embeddings	71.6
fastText embeddings (averaging out)	**80.9**

Based on the performance results, we decide to employ fastText embeddings as the main feature of our news model. The downgrade probability outputted from this is then combined with other quantitative measures in the benchmark, coupled with SMOTE oversampling technique, to build a final augmented model. Results of this final model are further explored in Section 4.

4 Final Model Results and Validation

4.1 Quantitative Evaluation

The following results are based on the evaluation on a 20% hold-out test set that are untouched, unseen throughout the training process. Both the benchmark and final models are tested on this same set to ensure fair comparison.

AUC. One of the main metrics we use to compare performances of different models is AUC, which tells how good a model is at distinguishing classes (the higher the number, the better). Figure 3 indicates that a downgrade model using news alone can achieve a test AUC of 80.9% and there is a 5% gain when adding this news probability on top of the existing model (from 82.7% for benchmark to 87.7% for final model).

Recall. Since the cost of a false negative in our business context (an impending downgrade event that the model is not able to capture) is very steep, we put an emphasis on optimizing the model's recall rate. Although a standalone news
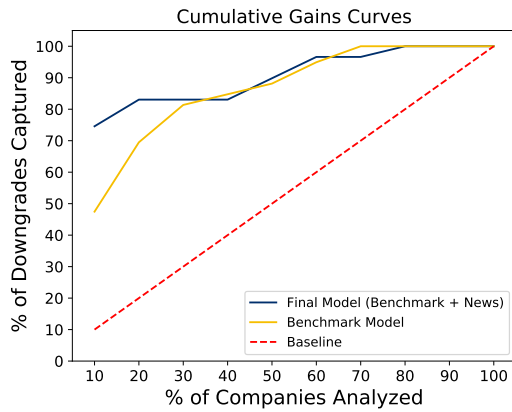
Figure 4: Cumulative gains chart. Given limited resources, the final model is able to provide a more optimal list of companies to analyze.

model produces a humble recall rate of 54.2%, incorporating news into benchmark model improves recall rate from 69.5% to 74.6% on the test set, resulting in a 5.1% gain in recall.

Cumulative Gains. Given limited human and time resources that can be dedicated to the task of analyzing companies in the list of anticipated downgrades, we would like to be able to capture as many downgrades as possible using as few test cases as possible. Figure 4 indicates that with news information added, analyzing the top 10% companies with the highest predicted downgrade probabilities can achieve a recall rate of 74.6%, which is a 27.1% improvement compared with the benchmark model. Similarly, analyzing the top 20% companies using the final model results in a recall rate of 83.1%, an increase of 13.6% compared with benchmark.

Overall, our model with the additional downgrade probability of news not only has higher accuracy rate (in terms of AUC and recall) but is also more efficient by providing a more optimal list of companies to analyze given restrained resources.

4.2 Qualitative Evaluation

To enhance the pragmatic sense of our model, we carefully examine all news articles associated with the true positives (actual downgrade events that the model is *able* to capture) and false negatives (actual downgrade events that the model is *unable* to capture) in our test set. Investigation shows that the model can pick up high-quality and very relevant article pieces to infer a downgrade event of a company.

True Positive Samples. There are 44 true positives corresponding with 16 unique companies in our test data. Following is a representative list of articles appearing in our true positive cases. Each example is shown at company level, contains only keywords due to space limit, and gives a good summary of the entire article piece.

- Company A: close factories; cut down 12,000 jobs; has been among the hardest to be hit by the trade war so far

- Company B: lose clients and get sued following misconduct revelations

- Company C: challenged by [an investor] for stretching itself financially to buy rival oil driller

- Company D: struggle with a host of issues; cut its dividend and report a wider-than-expected loss in its main engineering and construction unit

- Company E: hire restructuring firms and may choose to seek bankruptcy protection

- Company F: plan to wind down its dress-barn retail operations, resulting in the closure of about 650 stores

- Company G: anticipate having discussions on constructive basis relating to its underperformance

- Company H: make sophisticated missiles that use rare earth metals in their guidance systems, and sensors

- Company I: face lawsuit over art fraud; had been the willing auction house that knowingly and intentionally made the fraud possible

Although none of the articles explicitly mention that a company is being downgraded, the majority of them provide valuable insights into a company's financial health or how the company is perceived in the market. These signals could potentially be considered as leading indicators toward a downgrade. The model seems to be able to pick up subtle and indirect signs of an impending downgrade event as well. For example, in the case of Company H, the article refers to the company's heavy dependence on rare earth metals. This does not appear to relate to downgrade at first glance; however, more examination suggests that rare earth metal has become a political spotlight recently due to the trade tension between the U.S, who considers this mineral critical to the country's economic and national security, and China, who is the largest producer and manufacturer of this element in the world. Company H indeed was downgraded on 07-31-2019.

False Negative Samples. The test set includes 15 false negatives associated with 6 distinct companies. After review, we see these cases fall under either of two following categories:

- The news articles follow unpopular format structures compared with others in the dataset, which leads to their noisy quality even after preprocessing.

- The news articles appear in this false negative list are about companies that also exist in the true positive list. However, these pieces have a positive/neutral tone, were published shortly earlier, and might revolve around a different activity mentioned in the articles of the true positives. For example:

  – Company I's shareholders approve proposed acquisition; clients are pleased with the company's help to file their first lawsuit against the government (This piece was published on 06-21-2019. The one in true positive list was published on 06-25-2019. Company I's downgrade event happened on 09-17-2019).

  – Company C could divest most or all of [an investment vehicle] after buying out [another company]. (This
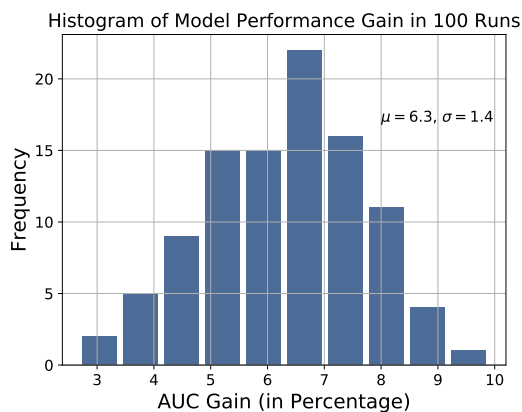
Figure 5: Histogram of the performance gain in 100 experiment runs. The gain is positive 100 times, with mean value 6.3% and standard error 1.6%.

piece was published on 06-14-2019. The one in true positive list was published on 06-27-2019. Company C's downgrade event happened on 08-01-2019).

Our takeaways here are that bad news can "outweigh" good/neutral news when it comes to contributing to the prediction of downgrade and companies' ratings could take downturn very shortly after emergence of some negative sentiments about them in the market.

4.3 Model Robustness

To evaluate the robustness of our final model's performance gain, we run 100 experiments using 100 different random seeds that are involved in the training/splitting, cross-validation and SMOTE oversampling processes. The performance gain in AUC from adding news to the benchmark model is positive 100 times out of 100 experiments. The mean value of these gains is 6.3% with a standard error of 1.6%, as shown in Figure 5.

This suggests that our positive performance gain is reliable and the 5% increase in AUC mentioned in Section 4.1 is still on the lower side. As we have access to a larger set of data, the standard error will theoretically be pushed closer to 0.

5 Conclusions and Future Work

In the hope of obtaining unique insights into companies' financial health that are not available in traditional quantitative credit measures, we train a downgrade risk model on solely news information. We demonstrate that news coverage, if represented appropriately in the data, can help detect adverse credit signal and considerably improve the performance of existing model that is trained on conventional credit measures.

There are multiple avenues for research opened from this work. Throughout our modeling development, we notice how we preprocess the text data could introduce a material change in our final model performance. Extracting only relevant information about the company of interest plays a

key role in increasing the gain. Thus, we would like to further explore more advanced techniques regarding information extraction, such as coreference resolution to find sentences that do not exactly include a company's name but still refer to the same entity, or sentence segmentation to pick out sentences when the boundary is ambiguous. In addition, because the news model is built solely based on online articles, it is inherently subject to media bias. Fact-checking news or tackling different forms of bias within mass media is outside the scope of this study but offers an exciting opportunity for future research.

6 Acknowledgement

References

Blei, D. M.; Ng, A. Y.; Jordan, M. I.; and Lafferty, J. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:2003.

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2016. Enriching word vectors with subword information. *CoRR* abs/1607.04606.

Hutto, C. J., and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Adar, E.; Resnick, P.; Choudhury, M. D.; Hogan, B.; and Oh, A. H., eds., *ICWSM*. The AAAI Press.

Le, Q., and Mikolov, T. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, II–1188–II–1196. JMLR.org.

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

Mikolov, T.; Grave, E.; Bojanowski, P.; Puhrsch, C.; and Joulin, A. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Nielsen, F. Å. 2011. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. *CoRR* abs/1103.2903.

Röder, M.; Both, A.; and Hinneburg, A. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, 399–408. New York, NY, USA: ACM.