Identifying Corporate Credit Risk Sentiments from Financial News

Noujoud Ahbali*, **Xinyuan Liu***, **Albert Aristotle Nanda***
Jamie Stark, Ashit Talukder, Rupinder Paul Khandpur
Moody's Analytics, 7 World Trade Center, New York, NY 10007, USA
{noujoud.ahbali, xinyuan.liu, albert.nanda
jamie.stark, ashit.talukder, rupinder.khandpur}@moodys.com

Abstract

Credit risk management is one central practice for financial institutions, and such practice helps them measure and understand the inherent risk within their portfolios. Historically, firms relied on the assessment of default probabilities and used the press as one tool to gather insights on the latest credit event developments of an entity. However, due to the deluge of the current news coverage for companies, analyzing news manually by financial experts is considered a highly laborious task. To this end, we propose a novel deep learning-powered approach to automate news analysis and credit adverse events detection to score the credit sentiment associated with a company. This paper showcases a complete system that leverages news extraction and data enrichment with targeted sentiment entity recognition to detect companies and text classification to identify credit events. We developed a custom scoring mechanism to provide the company's credit sentiment score (CSS^{TM}) based on these detected events. Additionally, using case studies, we illustrate how this score helps understand the company's credit profile and discriminates between defaulters and non-defaulters.

1 Introduction

Motivation. Historically, financial institutions performed credit risk management with techniques based on two different approaches (Chatterjee, 2015). The first approach is structural models, based on (Black and Scholes, 1973) and (Merton, 1974), which use the company's assets and liabilities to derive its probability of default. The second approach is default intensity models, also called reduced form models, developed by (Jarrow and Turnbull, 1995) and (Grundke and Riedel, 2004), which measure the default event as a statistical process, a random event following Poisson law, without considering the company's assets or liabilities.

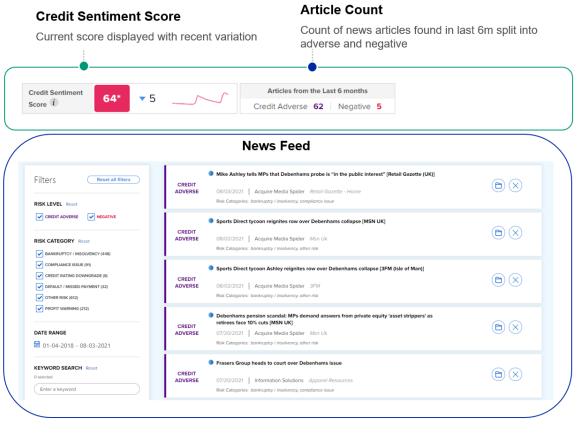

Figure 1: A screenshot of our deployed application.

These historic methods focus primarily on assessing the probability of default, which is useful in credit risk management. However, they are not designed to gain insights about a company's credit overall situation or identify the negative and credit adverse events the company has experienced or is likely to experience. This task falls under the responsibility of financial experts who may rely on news to identify such events, but this activity is considered highly tedious and time-consuming. Moreover, companies are increasingly covered in the press and journalists nowadays not only report facts, but go beyond in their analysis by making predictions, releasing warnings as well as establishing connections between companies.

Challenges. Most of the available news data is unannotated and un-exploitable at its initial state, which requires a significant entry effort for machine learning experiments. Furthermore, machine learning experiments in credit risk management has shown to boost accuracy in the default risk measure ('Oskarsd'ottir and Bravo, 2021) to show the effect of news sentiment on that same metric (Elena, 2020) or to focus on a single event prediction - credit downgrade in (Tran-The, 2020). However, none tackles news analysis automation and uses

*These authors contributed equally to this work

Proceedings of NAACL-HLT 2022: Industry Track Papers, pages 362 - 370
July 10-15, 2022

deep learning for credit event detection.

Our Goals. In order to derive explainable knowledge about a company's credit risk, we propose automating news analysis and identifying signals of negative and credit adverse events for companies. Such a method enables us to score the negative credit sentiment of companies. Our approach is a complete deployed application as shown in Figure 1. The enrichment pipeline starts with news collection (in English). It outputs a credit sentiment score (CSS) for companies based on the severity, recency, and volume of negative and credit adverse events detected from financial news articles. The custom Natural Language Processing (NLP) based pipeline's hallmarks include automated ingestion & filtering for finance-domain news articles, target-specific entity sentiment extraction. This pipeline allows high-precision content filtering and classification of the negative and credit adverse events mentioned in news articles (classified into five risk categories).

Our Contributions. The key contributions of this paper are:

- A novel, data-driven approach to detecting credit adverse events with targeted-entity sentiments
- A custom credit scoring methodology for companies from news, traditionally performed by financial experts.
- Extensive experimentation on real-world data on which our modeling approach performs well: including studies for defaulters VS non-defaulters and analysis of the discriminatory power of CSS between defaulters and non-defaulters.

2 Related Work

2.1 Aspect-Level Sentiment Analysis

When scientists prepare fine-grained sentiment models, they usually tackle the tasks of Aspect-based sentiment analysis (ABSA) (Do et al., 2019) and Targeted ABSA (TABSA) (Ma et al., 2018), where the latter considers the sentiment regarding a specific entity. Researchers have added context-dependencies to pretrained self-attention based language models called QACG-BERT (Wu and Ong, 2021) to improve the performance better. A mutual learning framework is used to take advantage of unlabeled data to assist the aspect-level sentiment-controllable review generation, consisting of a generator and a classifier that utilize confidence mechanism and reconstruction reward to enhance each other (Chen et al., 2021).

2.2 Deep Learning in Text Sentiment Analysis

A RNN model with LSTM units is trained based on Glove Embeddings of 400K words to predict the polarity (i.e., positive or negative sentiment) of the news (Souma et al., 2019). Moreover, an ensemble of CNN (Kim, 2014), LSTM (Hochreiter and Schmidhuber, 1997) and GRU (Chung et al., 2014) and a classical supervised model based on Support Vector Regression (SVR) is constructed which performs impressively on Microblog (Twitter and StockTwits) and news headlines datasets (Akhtar et al., 2017). Researchers have found that CNN is an effective model for predicting the sentiment of authors in the StockTwits dataset, among other models of logistic regression, doc2vec, and LSTM (Sohangir et al., 2018). A BERT model for the financial domain (FinBERT) pre-trained on a financial corpus and fine-tuned for sentiment analysis has shown promising results (Araci, 2019).

2.3 Machine Learning in Credit Risk

A study has shown that tree-based models are more stable than the models based on multilayer artificial neural networks in predicting loan default probability with structural features of financial conditions of a company (Addo et al., 2018). In addition, researchers have provided further evidence that regardless of the number of features used, boosted models outperform Linear Models, Decision Trees, and Neural Networks (Torrent et al., 2020). Further studies have stated that deep learning lends itself particularly well to analyzing textual data, but the improvement on numerical data is limited compared to traditional data mining models (Mai et al., 2019). Regarding Micro, Small, and Medium Enterprise (mSME) credit risk modeling, deep learning models, including the BERT model, appear to be robust concerning the quality of the text and therefore suitable for partly automating the mSME lending process because of their power to predict default based on textual assessments provided by a lender (Stevenson et al., 2021). In this study (Tran-The, 2020) a more NLP-focused approach is taken, using a combination of topic modeling and sentiment lexicons.

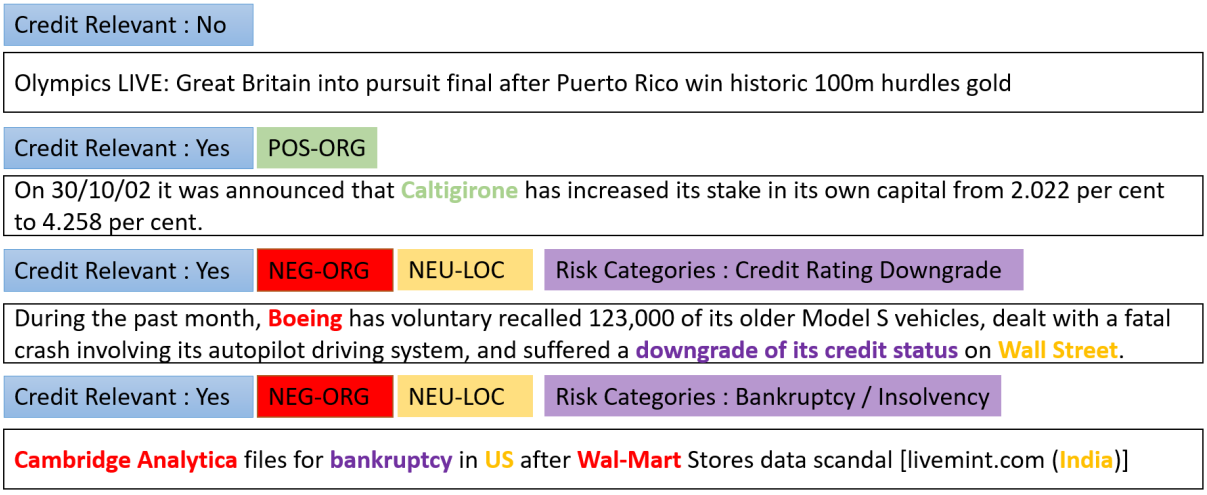

Figure 2: Annotated sentences by Credit Relevance, Target Entity Sentiment and Risk Categorization models.

3 Our Approach

In this section, we discuss different components of our scalable NLP pipeline that can ingest and infer English news from a news data source (Moody's Analytics NewsEdge [1]) that has over 170M articles, with an average of 150K news articles daily volume. To efficiently process large volumes of data, we have designed a data funnel process.

We have a credit relevance model at the head of the funnel, which helps discard irrelevant documents, viz. sports/technology-related articles. This model filters out 70% of the incoming documents. Next, the Target Entity Sentiment (TES) model extracts and tags all the entities in a document with Positive, Negative, and Neutral sentiment polarity, respectively. Following this step, the Risk Categorization model then classifies each sentence in the article into appropriate risk categories (discussed later in this section). In Figure 2 we illustrate different examples of sentences as annotated by each model.

3.1 News Enrichment Pipeline

In the pipeline, news articles are enriched with the output of the three following models.

Credit Relevance Model. We define credit-relevant news as any news story that contains business & finance-related topics which mention one or more corporate entities. We trained a binary relevance classification model using news data to identify relevant news. We leveraged the Reuters [2] news topics classification system, where we mapped Reuters codes into two classes: (1) in-domain (such as Merger/Acquisition, Sales, and promotions) and (2) out-of-domain (such asArt, Sports). In Table 1 we show the label distribution in both the train and test sets for the model.

After the text pre-processing (removal of Html links, numbers, and stop words removed), it was used with TF-IDF weighted features (Aizawa, 2003). Due to the train set size of over 30 Million articles, we chose a linear Support Vector Machine (SVM) model, trained with stochastic gradient descent (SGD) in out-of-core learning (Benczúr et al., 2018) setup.

Label	Train set	Test set
Relevant	13,323,062	3,291,751
Not Relevant	10,442,654	2,647,689
Total	23,765,716	5,939,440

Table 1: Distribution of annotated dataset for Credit Relevance model.

Target Entity Sentiment Model. The raw documents are first split into sentences using syntok [3] and then on each sentence a pre-trained WordPiece tokenizer (Schuster and Nakajima, 2012) is applied. Finally, each sentence is represented as $\{t_1, t_2, \ldots\}$ and the corresponding case tags $\{t_1^c, t_2^c, \ldots\}$. Token case tags used in the model are described in Table 2. Then given this sequence $\{t_1, t_2, \ldots\}$, we feed it to pre-trained Electra Base model[4] (Clark

[1] https://newsedge.com/
[2] https://liaison.reuters.com/tools/topic-codes
[3] https://github.com/fnl/syntok
[4] https://huggingface.co/google/electra-base-discriminator

Case Label	Description
AU	All letters in the token are upper-case
AL	All letters in the token are lower-case
IU	Only the initial letter of the token is upper-case
NU	All characters are digits(0-9)
MN	Most of the characters are digits
SN	Token has a digit

Table 2: Token case tags.

et al., 2020) to obtain contextual embeddings for each token $\{e_1, e_2, \ldots\}$. As shown in Figure 3, the contextualized embeddings are concantened with case embeddings $\{e^c_1, e^c_2, \ldots\}$ and fed to a linear layer to obtain the labels $\{\hat{y_1}, \hat{y_2}, \ldots\}$. To compute the loss, we used masked cross-entropy. And a dropout layer for regularization was added as well. The network was optimized using AdamW (Loshchilov and Hutter, 2019) optimizer.

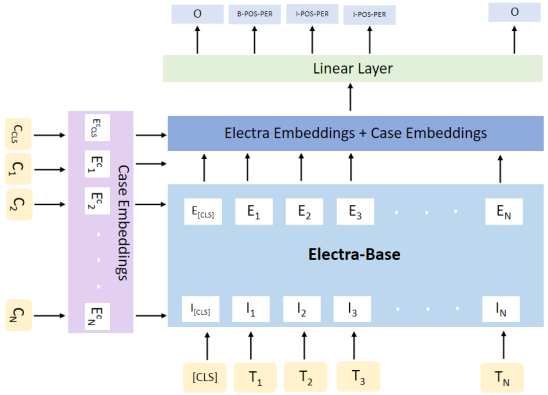

Figure 3: Architecture of Target Entity Sentiment Model.

To annotate data, each sentence was shown to 5 analysts, and those with majority consensus were selected; those with no clear majority were discarded. In Table 3 we show the overall distribution of labels across our final annotated dataset of 9,859 unique sentences, with an 80:20 split for training and evaluating the model.

Named Entity	Count	NEU	POS	NEG
PER	3585	67.92%	7.62%	24.46%
ORG	9020	63.47%	15.42%	21.11%
LOC	3824	92.89%	3.53%	3.58%
MONEY	2138	100%	0.00%	0.00%
MISC.	3020	92.29%	4.17%	3.54%

Table 3: Distribution of annotated dataset for Target Entity Sentiment model.

Risk Categorization Model. With similar preprocessing steps as inputs to the TES model, we trained a multi-label classification model, with an pre-trained Electra base model[5] (Clark et al., 2020), followed by convolutional layers (Kim, 2014) and a linear layer. We also used dropout to reduce overfitting and a sigmoid layer to generate the final prediction output. The final tuned hyperparameters for the model are listed in Table 7 in Appendix. We trained the model with different model architectures and hyperparameters over 30 epochs in each model training iteration and saved the best epoch on the test set. Our team of 4 annotators (among the paper's authors) was engaged in data labeling and cross-review activities for more than 60 hours to build the dataset. A tagging guideline was first discussed and agreed upon with explicit definitions of the seven labels. The risk categories labels and examples are listed in Table 4. Around 7000 sentences were collected and labeled according to the tagging guidelines to form the train and test sets (using stratified sampling). The distribution of labels in the train and test sets are listed in Table 8 in Appendix.

Credit Risk Scoring Model. Each company is scored daily using credit adverse news articles for the company, as tagged by Risk Categorization Model.

Step 1: For each date, we calculate the category weights w^{date}_{cat} over a fixed window of days. This is done by counting the number of articles in each category and using an exponential decay so more recent counts have more weights, as shown in Formula 1.

$$w^{date}_{cat} = \sum_{i=from}^{date} \text{count}^{i}_{cat} * e^{(date-i)/k} \quad (1)$$

where:

$$\begin{aligned} from &= \text{start date of the fixed window used for the calculations} \\ \text{count}^{i}_{cat} &= \text{count of all articles found for a given category (cat) on day (i)} \\ k &= \text{decay constant} \end{aligned}$$

Step 2: For each date, we calculate the category scores $score^{date}_{cat}$ by transforming the weights using a sigmoid function, which has the effect of capping the weight and also ensuring that only one or two articles mentioned will have limited impact.

[5] https://huggingface.co/google/electra-base-discriminator

Risk Category	Definitions	Example Sentences	Fixed Score
Bankruptcy / Insolvency	Proceedings of bankruptcy, insolvency or foreclosure, mentions of restructuring, administration or refinancing due to liquidity issues.	The British firm filed for Chapter 7 bankruptcy protection late Thursday.	100
Default / Missed Payments	Any mention of unpaid debts by a entity or the prospect of default for an entity.	Levitt home-building unit gets loan default notices.	75
Credit Rating Downgrade	Downgrades from rating organizations.	Standard Chartered's Shares Plunge 7% After Fitch Downgrade.	30
Profit Warning	Revenue, sales or EPS fall.	Carillion has been fighting for survival after contract delays and a drop in new business led to three profit warnings last year.	20
Compliance Issue	Any kind of financial crime, investigations, lawsuits, or violations.	TransAtlantic Petroleum Announces Notice of Noncompliance With NYSE MKT Continued Listing Standards.	2.5
Other Risk	Any type of company or credit relevant risks not covered in one of the five risk categories above.	On June 28, 2017, Southern Company and its subsidiary, Mississippi Power, suspended operations involving the coal gasifier portion of the Kemper County energy facility.	0
Not Relevant	Any text that is not evolved with credit risk.	Marks & Spencer to issue its first junk bond after reporting its first loss since joining the stock market in 1926.	0

Table 4: Risk Categories definitions with examples and weights in entity scoring

We multiply by a fixed score for that category as described in equation 2.

$$\text{score}_{cat}^{date} = \text{fixed}_{cat}/(1 + e^{-m*(w_{cat}^{date} - c)}) \quad (2)$$

where:

m = steepness of sigmoid function
c = number of articles needed to reach the midpoint of sigmoid function
fixed_{cat} = fixed score for a given risk category

The more severe the credit event is, the higher the fixed score is, as shown in Table 4.

Step 3: The Credit Sentiment Score at date t is the maximum category scores:

$$CSS^{date} = max(\text{score}_{cat}^{date}) \quad (3)$$

Our scoring function has an exponential decay which recognizes that news has a lasting value and impact during a specific period. It is reactive to the latest news as it weights recent news higher than older ones. The risk scores in the Credit Risk Scoring model are calculated via heuristics, as we do not have enough training data for a supervised approach. The fixed category score of each risk category in the Credit Risk Scoring model is shown in Table 4.

4 Evaluation

This section regroups the models evaluation as well as examples of case studies conducted on real-world data.

The Baseline. A simplified set of baseline models consists of three event relevance (binary classification) models instead of the multi-label classification model in the Risk Categorization model: Bankruptcy, Default, and Adverse News. This baseline method was actually in the earliest version of CSS^{TM} product, where we only considered the two most severe credit risk events: Bankruptcy and Default, in addition to a general class of less severe events called Adverse News. Bankruptcy and Default models handle bankruptcy and default-related events, respectively, while the Adverse News model deals with other credit events such as credit rating downgrade and illiquidity.

$$(C * m_t + \sum_{i=1}^{n} \text{article}_i^T)/(C + n) \quad (4)$$

where:

C = average number of articles per day in the last 10 days
m_t = historical daily score mean in last 10 days
n = number of articles in day t
article_i^T = i-th article score on day t

Each model outputs a score representing the prediction confidence about the underlying event from 0 to 100 for the input paragraph. The Bankruptcy and Default models are LSTM models (Hochreiter and Schmidhuber, 1997). And the Adverse News model is an LSTM model with attention mechanisms (Bahdanau et al., 2015) as these events are usually not as explicitly mentioned in the articles as the Bankruptcy and Default events.

These three classes are only present in the baseline, and we have added new risk categories (as shown in Table 4) for our latest version of the Risk Categorization Model. During the inference stage, each article is split into paragraphs fed to the three event relevance models. The paragraph score is the maximum score of the three relevance models, and the article score is the maximum score of all the paragraph scores within the article. Since Bankruptcy events are the most severe events while Adverse News are the least severe ones, we have applied weightings to the article scores of the three events with 100%, 75%, and 50%, respectively. At the company level, related articles are scored and are aggregated using a bayesian averaging, as shown in equation 4, to generate the company's daily sentiment score.

Models Evaluation. In Table 5 we show the classification report for the Credit Relevance Model on the test set (an overall F1-Score of 87%). As reported in Table 6 (detailed report in Table 10 in Appendix), the overall F1-Score of Target Sentiment Model on the test set is 77%, which shows best performance for ORG (Organization), the most relevant entity for our purpose. As reported in Table 6 (detailed report in Table 9 in Appendix), the overall F1-Score of Risk Categorization Model on the test set is 83%. We also notice better results for three of the four major credit events in the Credit Risk Scoring Model (Bankruptcy/ Insolvency, Credit Rating Downgrade, and Profit Warning).

	Precision	Recall	F1	Support
Not Relevant	86%	86%	86%	2647689
Relevant	89%	89%	**89%**	3291751

Table 5: Credit Relevance results.

	Precision	Recall	F1	Support
TES	76%	79%	77%	5391
RiskCat	83%	82%	83%	2146

Table 6: Performance (micro average) results for Targeted Entity Sentiment (TES) and Risk Categorization (RiskCat) model results.

The weights of risk categories in the Credit Risk Scoring model indicate the importance of the related credit events. That analyzes a company's creditworthiness. It coincides with the fact that we have better classification results in the Risk Categorization Model for the credit events that contribute with higher weights in the Credit Risk Scoring Model. As for Default / Missed Payments risk, its performance is close to the average performance.

To validate that our scoring model picks up credit adverse events for more than 6000 companies, we collect 40,000 negative articles over two years (2016 -2018) and corresponding default dates of defaulters. Of these companies, 1192 experienced a severe credit event (Bankruptcy/Insolvency or Default/Missed Payments), and the remaining became our control group. We refer to the former as defaulters and the latter as non-defaulters. We further filtered companies based on their newsworthiness to keep the ones with at least an article per month on average. In the end, the defaulters' group contains 1166 companies, whereas the non-defaulters have 3009 companies.

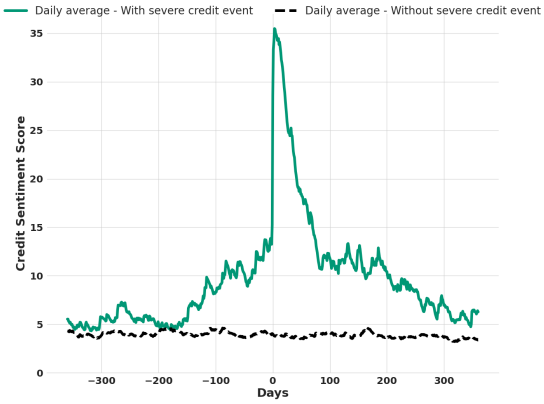

Figure 4: CSS Comparison between defaulters and non-defaulters

In Figure 4 we show the daily average CSS of the companies a year before and after the credit event (represented as the "0" date on X-axis). For comparison, we show the average score for the control group. The event dates for non-defaulters are chosen randomly during the same period as defaulters. The average CSS moves away from the long-term average towards the credit event. At

around three months before the credit event and until five months afterward, the score is around two times compared to the non-defaulters average. The peak of the defaulters after default events is around 35 after taking the average within the defaulters' group. However, not around 80 as in an individual company when a default event happens. Still, in Figure 4 we clearly distinguish defaulters and non-defaulters around default events by their average credit sentiment scores.

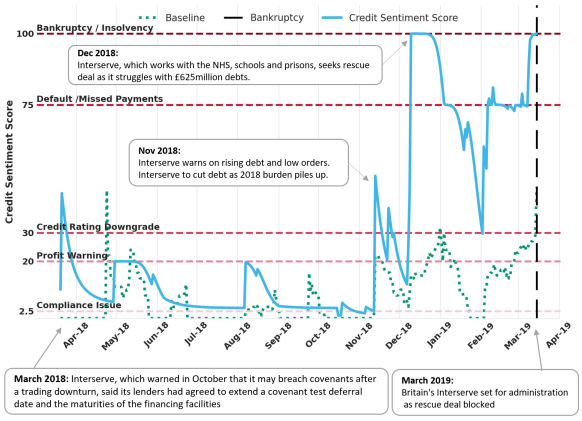

Figure 5: CSS and Baseline - INTERSERVE PLC.

Additionally, to validate the discriminatory power of CSS to identify the default and non-defaulting companies, we ran the following statistical tests. With *Kolmogorov–Smirnov* test (Massey Jr, 1951), we observed the Credit Sentiment Scores of the two groups (defaulters and non-defaulters) were statistically different, with a confidence level of 95%. Meanwhile, a *Mann–Whitney U* test (Nachar, 2008) showed that the probability of a defaulter's score is more significant than a non-defaulter's score (both selected randomly from the two groups) is statistically higher than 50%, with a confidence level of 95%.

Case Studies. To illustrate, we compared our CSS model to the baseline for defaulters and non-defaulters. As shown in Figure 5, CSS for Interserve PLC reacted to an early credit adverse signal (driven by Profit Warning and Default/Missed Payments) stronger compared to the baseline a year before the company was set for administration. Later, the news picked up a strong Bankruptcy / Insolvency signal as the company was seeking a rescue deal before it was set into administration.

In another example, in Figure 6 we show a consistently low CSS (as expected for the company as it is a non-defaulter company) compared to the baseline. This result is due to the baseline system

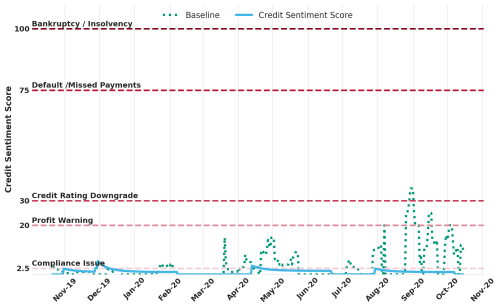

Figure 6: CSS and Baseline - AIR LEASE.

noise, as the articles often mention Air Lease's partners going into liquidation and insolvency issues. The result also shows that misclassifications on the paragraph level are way noisier than at the sentence level. A paragraph may have multiple sentences which refer to different companies with different sentiments in different contexts.

5 Conclusion

In this paper, we have designed, implemented, and deployed a deep-learning/NLP-powered application. This application can assist credit analysts in processing large amounts of news data and detecting and understanding the negative and credit averse events for companies. The pipeline utilizes various machine learning and deep learning models for data filtering, entity recognition sentiment analysis, and text classification.

As validated by the case studies and the modeling evaluation, the output sentiment score can distinguish between defaulted and non-defaulted companies. Since we only expose the CSS product instead of the complete models to the public, we will guarantee the truthiness of our in-house news source so that the system cannot be misused by publishing fake news.

We plan to use credit sentiment score as a signal to predict future credit events in future work. For example, given a company's credit sentiment score of a certain level, the probability that the target company will have some credit events within a certain period. We could also explore the sentiment analysis for positive credit events, aggregate company level scores into industry or region level, or focus on entities other than companies.

References

Peter Martey Addo, Dominique Guegan, and Bertrand Hassani. 2018. Credit risk analysis using machine and deep learning models. *Risks*, 6(2):38.

Akiko Aizawa. 2003. An information-theoretic perspective of tf–idf measures. *Information Processing Management*, 39(1):45–65.

Md Shad Akhtar, Abhishek Kumar, Deepanway Ghosal, Asif Ekbal, and Pushpak Bhattacharyya. 2017. A multilayer perceptron based ensemble technique for fine-grained financial sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 540–546, Copenhagen, Denmark. Association for Computational Linguistics.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *ArXiv preprint*, abs/1908.10063.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

András A. Benczúr, Levente Kocsis, and Róbert Pálovics. 2018. Online machine learning in big data streams.

Fischer Black and Myron Scholes. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–54.

Somnath Chatterjee. 2015. *Modelling credit risk*. Number 34 in Handbooks. Centre for Central Banking Studies, Bank of England.

Huimin Chen, Yankai Lin, Fanchao Qi, Jinyi Hu, Peng Li, Jie Zhou, and Maosong Sun. 2021. Aspect-level sentiment-controllable review generation with mutual learning framework. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12639–12647.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Hai Ha Do, PWC Prasad, Angelika Maag, and Abeer Alsadoon. 2019. Deep learning for aspect-based sentiment analysis: a comparative review. *Expert Systems with Applications*, 118:272–299.

Makeeva Elena. 2020. News sentiment in bankruptcy prediction models: Evidence from russian retail companies. , 14(4):7–18.

Peter Grundke and Karl O. Riedel. 2004. Pricing the risks of default: A note on madan and unal. *Review of Derivatives Research*, 7(2):169–173.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Robert Jarrow and Stuart M Turnbull. 1995. Pricing derivatives on financial securities subject to credit risk. *Journal of Finance*, 50(1):53–85.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yukun Ma, Haiyun Peng, and Erik Cambria. 2018. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5876–5883. AAAI Press.

Feng Mai, Shaonan Tian, Chihoon Lee, and Ling Ma. 2019. Deep learning models for bankruptcy prediction using textual disclosures. *European journal of operational research*, 274(2):743–758.

Frank J Massey Jr. 1951. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78.

Robert Merton. 1974. On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance*, 29(2):449–70.

Nadim Nachar. 2008. The mann-whitney u: A test for assessing whether two independent samples come from the same distribution. *Tutorials in Quantitative Methods for Psychology*, 4.

Mar'ıa 'Oskarsd'ottir and Cristi'an Bravo. 2021. Multilayer network analysis for improved credit risk prediction. *Omega*, 105:102520.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152. IEEE.

Sahar Sohangir, Dingding Wang, Anna Pomeranets, and Taghi M Khoshgoftaar. 2018. Big data: Deep learning for financial sentiment analysis. *Journal of Big Data*, 5(1):1–25.

Wataru Souma, Irena Vodenska, and Hideaki Aoyama. 2019. Enhanced news sentiment analysis using deep learning methods. *Journal of Computational Social Science*, 2(1):33–46.

Matthew Stevenson, Christophe Mues, and Cristi'an Bravo. 2021. The value of text for small business default prediction: A deep learning approach. *European Journal of Operational Research*.

Neus Llop Torrent, Giorgio Visani, and Enrico Bagli. 2020. Psd2 explainable ai model for credit scoring. *ArXiv preprint*, abs/2011.10367.

Tam Tran-The. 2020. Modeling institutional credit risk with financial news. *ArXiv preprint*, abs/2004.08204.

Zhengxuan Wu and Desmond C Ong. 2021. Context-guided bert for targeted aspect-based sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

6 Appendix

6.1 Risk Categories Model

Risk Categorization model final set of hyperparameters are shown in Table 7.

Hyperparameter	Value
Best Epochs	13
Max Length of Input Text	300
Train Batch Size	8
Test Batch Size	16
Initial Learning Rate*	1e-05
Dropout	0.7

* A learning rate scheduler is implemented to decrease the learning rate in later epochs to better converge and reduce overfitting

Table 7: Tuned Hyperparameters in Risk Categorization Model.

Risk Category	Train set	Test set
Profit Warning	688	329
Bankruptcy / Insolvency	853	372
Compliance Issue	326	161
Default / Missed Payments	596	309
Credit Rating Downgrade	426	204
Other Risk	1347	544
Not Relevant	596	227
Total	4832	2146

Table 8: Distribution of annotated dataset for Risk Categories model.

In Table 8 we show the distribution of the model's annotated dataset (train and test sets) along with the detailed classification report on the test set in Table 9.

Labels	Precision	Recall	F1-Score	Support
Profit Warning	**86%**	**89%**	**87%**	**329**
Bankruptcy / Insolvency	**93%**	**94%**	**94%**	**372**
Compliance Issue	81%	60%	69%	161
Default / Missed Payment	**79%**	**83%**	**81%**	**309**
Credit Rating Downgrade	**95%**	**95%**	**95%**	**204**
Other Risk	75%	76%	75%	544
Not Relevant	79%	68%	73%	227
Micro Avg	83%	82%	83%	2146
Macro Avg	84%	81%	82%	2146

Table 9: Detailed Risk Categories results.

6.2 Target Entity Sentiment Model

The primary entity label that our pipeline relies on is - *Organization* (Org) as our focus is on corporate entities. Accurately discerning the sentiment polarity (Pos, Neg, Neu) of these target Organizations is an essential requirement of the pipeline, and in the Table 10 we highlight the F1 score on these three classes.

Entity Type	Precision	Recall	F1-Score	Support
Money	94%	96%	95%	502
Neg Loc	57%	32%	41%	25
Neg Misc	36%	30%	33%	30
Neg Org	**66%**	**70%**	**68%**	**514**
Neg Per	72%	67%	69%	220
Neu Loc	85%	89%	87%	890
Neu Misc	71%	76%	73%	676
Neu Org	**74%**	**79%**	**77%**	**1612**
Neu Per	78%	80%	79%	518
Pos Loc	46%	24%	32%	25
Pos Misc	44%	44%	44%	27
Pos Org	**66%**	**69%**	**67%**	**298**
Pos Per	54%	70%	61%	54
Micro Avg	76%	79%	77%	5391
Macro Avg	65%	64%	64%	5391

Table 10: Detailed Targeted Sentiment results.