Identifying Corporate Credit Risk Sentiments from Financial News

Anonymous ACL submission

Abstract

Credit risk management is one major practice for financial institutions, that helps them measure and understand the inherent risk within their portfolios. Historically, they relied on the assessment of default probabilities (via structural or default intensity models) and used the press as one tool to gather insights on the latest credit event developments of an entity. However, because the current news volume and coverage for companies is generally heavy, 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, with the aim of scoring the credit sentiment associated with a company in order to assist credit risk management efficiently. The result is a complete system leveraging news extraction and data enrichment (with targeted sentiment entity recognition to detect companies and text classification to identify credit events), as well as a custom scoring mechanism designed to provide the company's credit sentiment, called Credit Sentiment Score™ (CSS). Additionally, studies are shown to illustrate how CSS helps to gain knowledge about the company's credit profile but also discriminates between defaulters and non-defaulters.

1 Introduction

One of the biggest challenges for financial institutions today is to assess, manage and mitigate the credit risk inherent in loan and investment portfolios. Other than just for meeting regulatory requirements (established after the 2008 financial crisis), properly assessing and managing credit risk can also reduce the severity of losses.

Motivation. Historically, financial institutions have been tackling this problem internally or using third party frameworks 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. Structural models provide an intuitive and economic explanation of the default, however, they require full information of the company's balance sheet and assume that the assets value is observed, while in fact, it is not. 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. While such a method requires less detailed knowledge about the company's assets and liabilities compared to structural models, no economical meaning is attached to the default, which makes such modelling lack explain-ability in the real world. 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 as a highly tedious and time-consuming task. Indeed, companies are increasingly covered in the press and journalists not only report facts, but go beyond in their analysis by making predictions, releasing warnings as well as establishing connections between companies. Indeed, news stories help into shaping an instant image of the company's current situation, which makes them a valuable source in understanding the company's credit profile.

Challenges. Most of the available news data is un-annotated and un-exploitable at its initial state, which requires a significant entry effort for machine learning experiments. And even if the recent developments of Natural Language Processing

(NLP) techniques and computational power (Torfi et al., 2021) enhance machines' ability to extract value from the human language, domain-specific language modelling is essential to boost performance (Coden et al., 2005). Furthermore, machine learning experiments in credit risk management were conducted 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). But we found none of them tackles news analysis automation and deep-learning powered credit event detection.

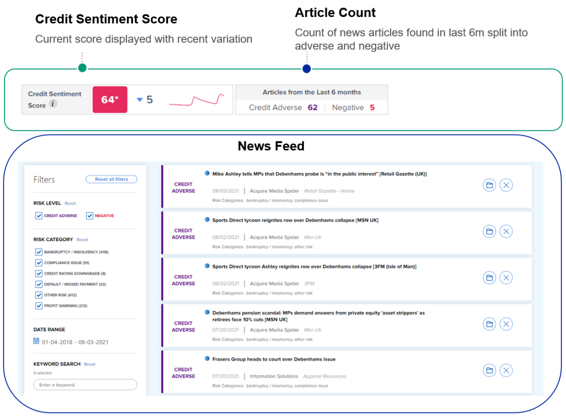

Figure 1: An screenshot of our Credit Sentiment Score system.

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. This enables us to score the negative credit sentiment of companies. Our approach is a complete deployed application as shown in 1. The enrichment pipeline that starts with news collection (in English) and outputs a Credit Sentiment Score (CSS) for companies, on the basis of the severity, recency and volume of negative and credit adverse events detected from financial news articles. The hallmarks of the custom Natural Language Processing (NLP) based pipeline include - automated ingestion & filtering for finance-domain news articles, target-specific entity sentiment extraction that allows high-precision content filtering and classification the negative and credit adverse events mentioned in news articles are classified in 5 risk categories.

Our Contributions. The key contributions of this paper are:

- A novel, deep learning powered approach to analyze and detect credit adverse events from news at sentence and entity levels, broken down into Natural Language Processing (NLP) tasks, which are traditionally performed by financial experts.
- A custom credit scoring methodology for companies based on credit adverse events.
- Extensive experimentation conducted on real world data on which our modelling approach performs well, including case studies for financially distressed companies and analysis of the discriminatory power of CSS between defaulters and non-defaulters.

2 Related Work

2.1 Aspect-Level Sentiment Analysis

When people are doing fine-grained sentiment models, they usually tackle the tasks of Aspect-based sentiment analysis (ABSA) and Targeted ABSA (TABSA), where the latter considers the sentiment regarding specific entity. Researchers have added context-dependencies to pretrained self-attention based language models called QACG-BERT (Wu and Ong, 2020) to better improve the performance. A mutual learning framework is also brought up to take advantage of unlabeled data to assist the aspect-level sentiment-controllable review generation, which consists of a generator and a classifier which utilize confidence mechanism and reconstruction reward to enhance each other (Chen et al., 2021). We have utilized our fine-grained NER model to tackle target entity recognition and sentiment analysis in a multi-task learning setting that combines sentence and entity-level contexts. This model is used in our pipeline to capture the sentiment of target entities more accurately by a network of ELECTRA along with couple of task specific components trained on a dataset with labels of entity type and its sentiment for each entity in the sentence.

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, LSTM and GRU 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). Another paper proposes a new sentiment analysis model-SLCABG, which is based on the sentiment lexicon and combines CNN and attention-based Bidirectional Gated Recurrent Unit (BiGRU) on the book reviews data (Yang et al., 2020). This coincides with our risk category model that uses Electra base model followed by CNN layers to predict the events of financial news sentences, also our model focuses on credit related news instead of stock related news that people have done many works on.

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, people 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 with traditional data mining models (Mai et al., 2019). Regarding Micro, Small and Medium Enterprise (mSME) credit risk modelling, deep learning models including the BERT model appear to be robust to the quality of the text and therefore suitable for partly automating the mSME lending process because of its 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 topic modeling and sentiment lexicons (Tran-The, 2020).

3 Our Approach

In this section we discuss different components of our scalable NLP pipeline that can ingest and infer from a news data source (Acquire Media NewsEdge [1]) that has over 170M articles, with an of average of 500K news articles daily volume. To efficiently process large volumes of data, we have designed a data funnel process. At the head of the funnel, we have credit relevance model, this is a very fast(inference time) model which can process large number of documents very quickly. The idea here is to discard irrelevant documents viz. sports/technology related articles. This model filters out 70% of the incoming documents. Next in the funnel is Target Entity Sentiment(TES) model, this model tags all the entities in a document with Positive, Negative and Neutral tags per sentence. Following the TES model, we pass the documents through the Risk Categories Model(sentences which have been tagged by TES model). The annotated document is then saved in a Elastic Search DB for faster access and retrieval. The scoring function then works on the annotated articles with a user-specified date range.

3.1 News Enrichment Pipeline

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

Credit Relevance. Retaining only financially related news articles helps to remove irrelevant articles and reduce the input volume. For this purpose, we used a binary classification model, called Credit Relevance Model. The raw text is tokenized into tokens using TF-IDF vectorization (Aizawa, 2003), then fed to a linear Support Vector Machine (SVM) model, trained with stochastic gradient descent (SGD) out-of-core learning (Benczúr et al., 2018) using Scikit-learn (Pedregosa et al., 2011).

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 data set for Credit Relevance model.

Relevant news for Credit Relevance Model is by definition a financially related text that potentially holds information and knowledge about a company. To build data sets for training and testing the model, we relied on in-domain (such as Merger/Acquisition, Sales and promotions...) and out-of domain (Art, Sports ...) topics derived from news classification in Reuters[2]. Those domains

[1] https://newsedge.com/

[2] https://liaison.reuters.com/tools/topic-codes

were then mapped to (1) Relevant and (2) N
Relevant. Table 1 gives the label counts in both t
train and test sets.

Target Entity Sentiment Model. The raw d
uments are tokenized into sentences using syntok
and then each sentence is tokenized using a p
trained WordPiece tokenizer (Schuster and Nak
jima, 2012). Case information is also added to
the tokens as shown in Table 2

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-c
NU	All characters are digits(0-9)
MN	Most of the characters are digits
SN	Token has a digit

Table 2: Token case tags.

Finally, each sentence is represented as $\{t_1, t_2, \ldots\}$ and the corresponding case tags $\{t_1^c, t_2^c, \ldots\}$. The model architecture is shown in Figure 2. Given the tokens of a sentence $\{t_1, t_2, \ldots\}$, we feed it to pre-trained Electra Base model (Clark et al., 2020) to obtain contextual Electra embeddings for each token $\{e_1, e_2, \ldots\}$. The contextualized embeddings are concantened with case embeddings $\{e_1^c, e_2^c, \ldots\}$ 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:

$$L = \frac{1}{n} \sum_{i}^{n} l(\hat{y}_i, y_i) \tag{1}$$

where:

$l(\cdot, \cdot)$ = the cross-entropy function.

We also added dropout layer for regularization. The network was optimized using AdamW (Loshchilov and Hutter, 2017) optimizer.

Our team of machine learning scientists and financial experts works closely together to examine and finalize entity type and sentiment category. To collect sentiment labels, financial analysts were instructed to (1) select the entity specific sentiment (2) indicate their level of confidence. Each sentence was shown to 5 analysts and those with majority consensus were selected. Sentences with no clear majority were removed from the final dataset. This resulted in 9,859 out of a total 10,516 sentences.

[3] https://github.com/fnl/syntok

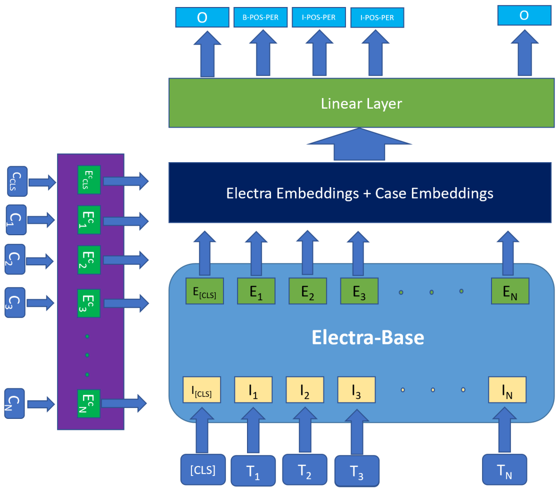

Figure 2: Targeted Sentiment Model architecture.

Our dataset has 4+1 named entity categories as described in Table 3. We did a 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 data set for Target Entity Sentiment model.

Risk Categories Model. In the same way as Target Entity Sentiment Model, sentences are tokenized. They are then fed to a multi-label classification model, which consists of a pre-trained Electra base model, 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. Apart from giving more granularity of the output and being at sentence level, this model benefits from a maintainable architecture, to add more labels (risk categories) for instance as well as re-training on other multi-classification tasks.

To build the Risk Categories Model, a team of 4 annotators (financial domain experts) were engaged in data labeling and cross-review activities for more than 60 hours. 7000 sentences were collected and labeled according to explicit and clear label definitions, to form the train and test sets (using stratified sampling). The distribution of labels in the train and test sets are listed in the below table 4. Figure 3 shows examples of sentences as anno-

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 4: Distribution of annotated data set for Risk Categories model.

tated by Credit Relevance, Target Entity Sentiment and Risk Categories models.

The baseline The baseline consists of three event relevance (binary classification) models: Bankruptcy, Default and Bad News. Each model outputs a score which is 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) while Bad News model is a LSTM model with attention mechanisms (Bahdanau et al., 2014). During the inference stage, each article is split into paragraphs which are 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 paragraphs scores within the article. Since Bankruptcy events are the most severe events while Bad News are the least severe ones, we have applied weightings on 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 Average to generate the company's daily sentiment score, as given by 2.

$$(C * m_t + \sum_{i=1}^{n} article_i^T)/(C + n) \qquad (2)$$

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

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

Step 1: For each date, calculate the category weights w_{cat}^{date} 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. i.e.:

$$w_{cat}^{date} = \sum_{i=from}^{date} count_{cat}^{i} * e^{(date-i)/k} \qquad (3)$$

where:

$from$ = start date of the fixed window used for the calculations
$count_{cat}^{i}$ = count of all articles found for a given category (cat) on day (i)
k = decay constant

Step 2: For each date, calculate the category scores $score_{cat}^{date}$. This is done by transforming the weights using a sigmoid function. This has the effect of capping the weight and also ensuring that only one or two article mentions will have limited impact. We then multiply by a fixed score for that category, i.e.:

$$score_{cat}^{date} = fixed_{cat}/(1 + e^{-m*(w_{cat}^{date} - c)}) \qquad (4)$$

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 5.

Risk Category	Fixed Score
Profit Warning	20
Bankruptcy/ Insolvency	100
Compliance Issue	20
Default / Missed Payments	75
Credit Rating Downgrade	30
Other Risk	0
Not Relevant	0

Table 5: Weighing scheme for the risk categories.

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

$$CSS^{date} = max(score_{cat}^{date}) \qquad (5)$$

As opposed to baseline, our CSS approach has an exponential decay which recognizes that news have a lasting value and impact during a certain period. It is reactive to the latest news as it weights recent news higher than older news.

Credit Relevant : No

Olympics LIVE: Great Britain into pursuit final after Puerto Rico win historic 100m hurdles gold

Credit Relevant : Yes | POS-ORG

On 30/10/02 it was announced that Caltigirone has increased its stake in its own capital from 2.022 per cent to 4.258 per cent.

Credit Relevant : Yes | NEG-ORG | NEU-LOC | Risk Categories : Credit Rating Downgrade

During the past month, Boeing has voluntary recalled 123,000 of its older Model S vehicles, dealt with a fatal crash involving its autopilot driving system, and suffered a downgrade of its credit status on Wall Street.

Credit Relevant : Yes | NEG-ORG | NEU-LOC | Risk Categories : Bankruptcy / Insolvency

Cambridge Analytica files for bankruptcy in US after Wal-Mart Stores data scandal [livemint.com (India)]

Figure 3: Annotated sentences by Credit Relevance, Target Entity Sentiment and Risk Categories models.

4 Evaluation

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

Models Evaluation. Table 6 shows the classification report for the Credit Relevance Model on the test set (an overall F1-Score of 87%) As reported in

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

Table 6: Credit Relevance results.

Table 7, the overall F1-Score of Target Sentiment Model on the test set is 77%. And we have decent prediction power regarding the extraction and sentiment for ORG(Organization), the most relevant entity for our purpose. As reported in Table 8, the overall F1-Score of Risk Categories Model on the test set is 83%. We also notice better results for the credit events that contribute with higher weights in the Credit Risk Scoring Model (as shown table 5) : Bankruptcy/ Insolvency, Credit Rating Downgrade and Profit Warning. As for Default / Missed Payments risk, its performance is close to the average performance. A validation study was performed to confirm that the Credit Risk Scoring Model picks up credit adverse events. For more than 6000 companies in total, over 40,000 negative articles were collected during a one-year period between 2016 -2018. Of these companies, 1192 experienced a severe credit event (bankruptcy, default, distressed exchange offer, etc.) and the remaining became our control group. We refer to the former as de-

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 7: Targeted Sentiment results.

faulters, and the latter as non-defaulters.We further filtered companies based on their news worthiness to keep the ones with at least an article per month on average.

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

Table 8: Risk Categories results.

Figure 4 shows the daily average CSS of the companies before and after the credit event (represented

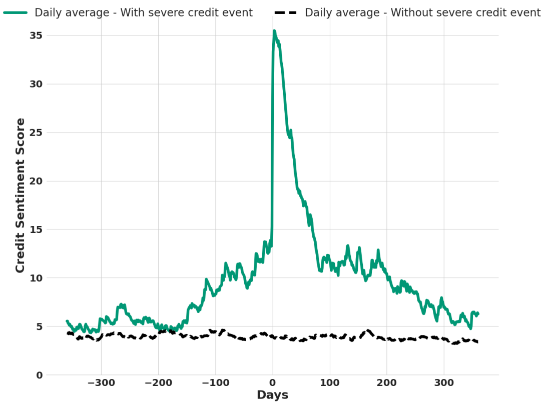

Figure 4: CSS Comparison between defaulters and non-defaulters

as the "0" date in X-axis). For comparison, the average score for the control group is shown.The average CSS moves away from the long-term average as it moves towards the credit event. At around three months before the credit event and until five months afterwards, the score is around two times the non-defaulters average.

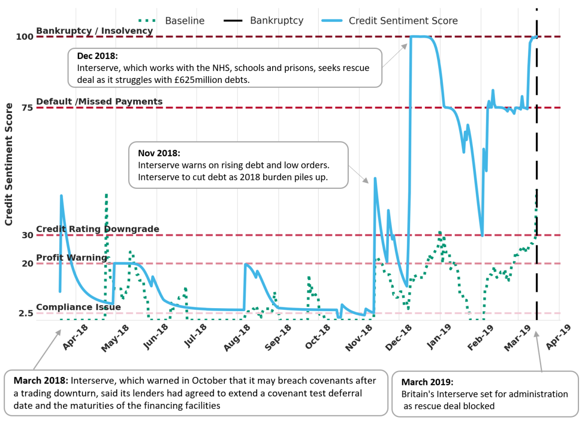

Figure 5: CSS and Baseline - INTERSERVE PLC.

Additionally, in order 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 (Jr., 1951), we observed that the Credit Sentiment Scores of the two groups (defaulters and non-defaulters) were statistically different, with a confidence level of 95%. And a *Mann–Whitney U* test (Nachar, 2008) proved that the probability of a defaulter's score being greater that a non-defaulter's score (both selected randomly from the two groups) is statistically higher than 50%, with a confidence level of 95%.

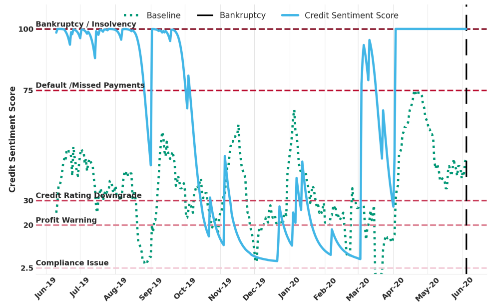

Figure 6: CSS and Baseline - DEBENHAMS PLC.

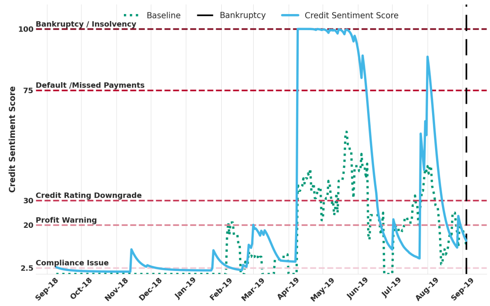

Figure 7: CSS and Baseline - SENVION SA.

Case Studies This section regroups examples of case studies to illustrate CSS compared to the baseline for defaulters and non-defaulters.

In Figure 5, CSS for Interserve PLC reacted to an early credit adverse signal (driven by Profit Warning and Default / Missed payment) stronger compared to the baseline, a year before the company was set for administration. Later, a strong Bankruptcy / Insolvency signal was picked up by the news as the company went into more severe credit events (seeking a rescue deal), before it was set into administration. Figure 6 shows the case of Debenhams PLC, for which a Bankruptcy / Insolvency signal was picked up, a year before the the company declared bankruptcy. The same is true in figure 7 for Senvion SA, which shows a strong Bankruptcy / Insolvency signal 6 months before the event itself.

In the other hand, Figure 8 shows a consistently low CSS (as expected for the company as it is a non-defaulter company), compared to the baseline. This is due to the baseline system noise, as the articles often mention Air Lease's partners going into liquidation and insolvency issues. This also shows that mis-classifications on the paragraph level are way noisier than the sentence level. Indeed, a para-

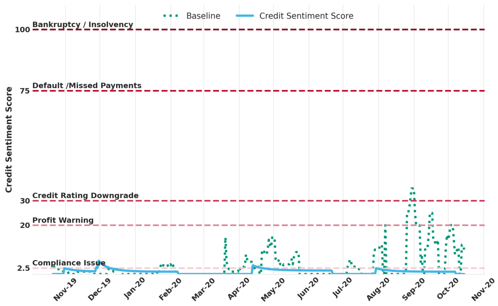

Figure 8: CSS and Baseline - AIR LEASE.

graph may have multiple sentences which refer to different companies with different sentiments in different contexts.

5 Conclusion

In this paper, we have designed and implemented a natural language processing pipeline that is capable of assisting credit analysts to process large amounts of news data, detect and understand the negative and credit averse events for companies. The pipeline utilizes various machine learning and deep learning models for data filtering, named entity recognition sentiment analysis and text classification. The output sentiment score is able to distinguish between defaulted and non-defaulted companies, as validated by the case studies and the modelling evaluation. In future work, we could explore the sentiment analysis for positive credit events, take other factors such us the industry or region into account 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.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

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.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Anni R. Coden, Serguei V. Pakhomov, Rie K. Ando, Patrick H. Duffy, and Christopher G. Chute. 2005. Domain-specific language models and lexicons for tagging. *Journal of Biomedical Informatics*, 38(6):422–430.

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.

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

Yoon Kim. 2014. Convolutional neural networks for sentence classification.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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.

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.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

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*.

Amirsina Torfi, Rouzbeh A. Shirvani, Yaser Keneshloo, Nader Tavaf, and Edward A. Fox. 2021. Natural language processing advancements by deep learning: A survey.

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

Tam Tran-The. 2020. Modeling institutional credit risk with financial news. *arXiv preprint arXiv:2004.08204*.

Zhengxuan Wu and Desmond C. Ong. 2020. Context-guided BERT for targeted aspect-based sentiment analysis. *CoRR*, abs/2010.07523.

Li Yang, Ying Li, Jin Wang, and R Simon Sherratt. 2020. Sentiment analysis for e-commerce product reviews in chinese based on sentiment lexicon and deep learning. *IEEE Access*, 8:23522–23530.