Finance and Economics Discussion Series

Federal Reserve Board, Washington, D.C.
ISSN 1936-2854 (Print)
ISSN 2767-3898 (Online)

Integrating Prediction and Attribution to Classify News

Nelson P. Rayl, Nitish R. Sinha

2022-042

Please cite this paper as:
Rayl, Nelson P., and Nitish R. Sinha (2022). "Integrating Prediction and Attribution to Classify News," Finance and Economics Discussion Series 2022-042. Washington: Board of Governors of the Federal Reserve System, https://doi.org/10.17016/FEDS.2022.042.

Integrating Prediction and Attribution to Classify News*

Nelson P. Rayl
Nitish R. Sinha

Draft: June 9, 2022

Abstract

Recent modeling developments have created tradeoffs between attribution-based models, models that rely on causal relationships, and "pure prediction models" such as neural networks. While forecasters have historically favored one technology or the other based on comfort or loyalty to a particular paradigm, in domains with many observations and predictors such as textual analysis, the tradeoffs between attribution and prediction have become too large to ignore. We document these tradeoffs in the context of relabeling 27 million Thomson Reuters news articles published between 1996 and 2021 as debt-related or non-debt related. Articles in our dataset were labeled by journalists at the time of publication, but these labels may be inconsistent as labeling standards and the relation between text and label has changed over time. We propose a method for identifying and correcting inconsistent labeling that combines attribution and pure prediction methods and is applicable to any domain with human-labeled data. Implementing our proposed labeling solution returns a debt-related news dataset with 54% more observations than if the original journalist labels had been used and 31% more observation than if our solution had been implemented using attribution-based methods only.

JEL codes: C40, C45, C55.

Keywords: News, Text Analysis, Debt, Labeling, Supervised Learning, DMR

* Our views do not necessarily reflect those of the Federal Reserve System nor its Board of Governors. Rayl is grateful to Matt Taddy. Sinha is grateful to Sameer Agarwal at Google for relating Efron's paper. Sinha (Nitish.R.Sinha@frb.gov) is at the Research and Statistics division of the Federal Reserve Board, 20th Street and Constitution Avenue, NW, Washington DC 20551; Rayl (nelsonrayl14@gmail.com) co-authored while he was a research assistant at the Board.

1. Introduction

In academic literature and among practitioners there is a divide between traditional (attribution-based) statistical techniques and what Efron (2020) calls pure prediction techniques (also sometimes pejoratively called black boxes).[1] Historically, traditional statistical techniques such as linear regressions have belonged to the fields of economics and other social sciences while pure prediction techniques such as random forests and neural networks have belonged to the fields of data science and computer science. Recent work such as Athey et. al. (2021) has called for greater consideration of how these differing technologies which have "historically been regarded as standing in opposition to one another" may be carefully synthesized to make better predictions, but thus far researchers have tended to favor one technology or the other. While preference for attribution techniques or prediction techniques is often driven by taste or loyalty to the methods of a given field, the differences between them are material. Traditional statistical techniques estimate the effect of candidate variables on an outcome of interest, and as a result a forecast created using these techniques has the benefit of being attributed to candidate variables. In contrast, pure prediction techniques forego attribution in favor of maximizing forecast accuracy. By utilizing a massive number of parameters or non-linear functions, pure prediction technologies sacrifice the ability to attribute the forecast to candidate variables, but in return promise greater prediction accuracy.[2] The differing objectives and methods of these two modeling technologies present tradeoffs between attribution and prediction that are rarely considered by practitioners of either paradigm. Practitioners of attribution-based methods have historically focused on the explanatory power of a model and ignored potentially useful datasets

[1] See table 5 of Efron (2020) for a summary of his terminology.

[2] While it is possible to directly map inputs to output in many pure prediction models, it is the computation and memory constraints of humans that prevent these models from being interpretable or attributable.

that have too many predictors to fit in their model. On the other side, much of the pure prediction and machine learning literature focuses on proving that methods have a high level of accuracy on a test set, sometimes without considering the fact that simple models would do just as well (see Athey et. al. for an example of flu prediction). We compare performance across these two different technologies and attempt to merge the benefits from both technologies to estimate label inconsistency in a large text-dataset. Our findings provide guidance for all domains in which the tradeoffs between attribution and prediction are material as well as a blueprint for how these historically isolated methods may be combined to improve prediction.

Text Analysis as a Domain for Attribution and Prediction

One domain in which neither attribution nor prediction dominates and the tradeoffs between these methods may be readily observed is text analysis. A brief scan of the literature reveals the usage of methods ranging from simple dictionary-based approaches where the occurrence of specific words are counted to the usage of transformer models with hundreds of billions of parameters (See Brown et al. (2020) for an example of one of these models). The diversity of models and methods used in textual analysis is a result of the richness of textual data itself. The ability to format text as individual words, pairs of words, or strings of text combined with the diversity and abundance of text gives researchers many candidate variables to build models around. As a result, text models have been developed that range the spectrum of interpretability, complexity, and performance, making text analysis an ideal domain for testing the tradeoffs between prediction and attribution. Textual data has also proven to be a compelling source of information for research and forecasting (See Gentzkow, Kelly and Taddy (2019) to get a sense of interest in text as a data source in economics). For example, Baker, Blume, and Stevens (2016) find mention of the word “uncertainty” to be a good predictor for a whole host of economic outcomes. While text data is a compelling predictor, the features that make it a rich data source for model development and forecasting also present unique problems that must be addressed. For one, the sheer size of textual data makes manual inspection of the data uneconomical. Another problem is that unlike experimental data where the data is created as the result of an experiment, text data is rarely a result of a targeted experiment, instead it is a by-product of people going about living their lives. For example, a common source of text data is archived newspapers or journals that contain articles on many subjects. A researcher using one of these datasets may want to consider only finance-related text, but without labels specifying each article as finance-related or non-finance-related they cannot verify that they are picking up finance-related information instead of information on sports or politics. To address this issue many text datasets come with labels that allow a researcher to filter the dataset to their desired information. This introduces another problem with textual data – inconsistent labeling and concept drift.[3] The labels used to classify text may change over time with the introduction of new labels or changes in labeling standards. If labeling standards are consistent over time, there is still the risk of concept drift as the association between text and label may change over time. For example, prior to 2008 an article about housing supply may not have been labeled as finance-related, but after 2008 the very same article may be considered as relevant to financial markets. A forecaster may want to include the pre-2008 article on housing supply to their dataset, but if concept drift is not accounted for this article would be left out. Even if a forecaster suspects that a textual dataset is consistently labeled and does not suffer from concept drift, this

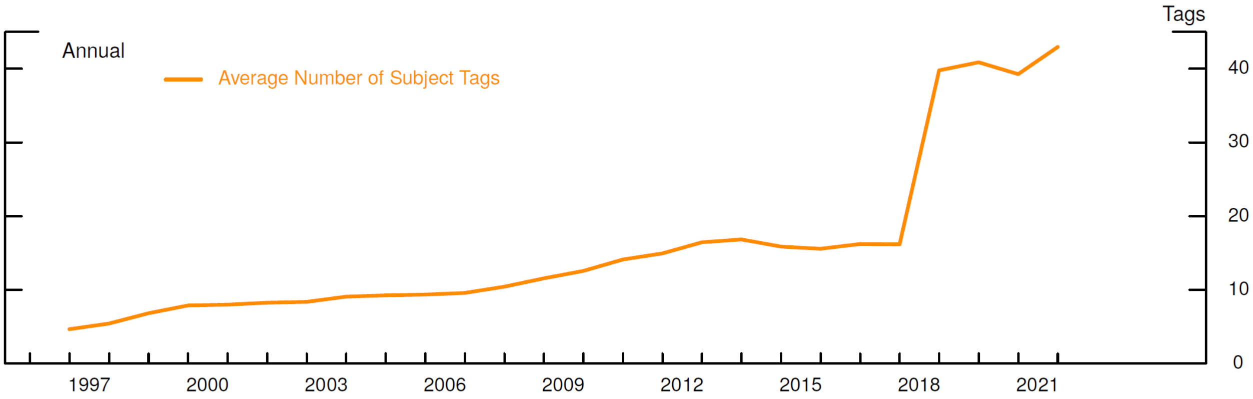
[3] Concept drift occurs when the statistical properties of a target variable change over time. In this case, we are considering changes in the properties that map the information in text to a specific subject label. The concept drift tagging literature has historically tried to infer online concept drift, i.e. to provide tags that are consistent with standards at the time of concept emergence rather than retrospective tagging which is typically important to macroeconomists and political scientists. See Helmbold and Long (1994) for an early example of concept-drift in machine learning literature.

cannot be verified without reading all of the text. The problem of inconsistent labeling also occurs most frequently in the domains in which textual data is most useful. In macroeconomics and politics where patterns occur over long time horizons and numeric data that goes back in history is nonexistent, textual data offers forecasters and researchers a valuable tool. In these domains there is an abundance of historical text that is ready to be digitized, however, this textual data is likely to face labeling issues that must be addressed to render the data useful.

Thomson Reuters News Archive

To make the discussion of labeling issues concrete, we present the case of the Thomson Reuters news archive which contains 27 million news articles published between 1996 and 2021. The archive contains all news broadcasted on the Reuters newswire service and is available for research, but it was not created for researchers. Instead, the archive is intended for traders and analysts who are interested in financial markets. The news articles contained in this dataset range in subject from finance to politics and the arts, and each news article is labeled with multiple subject tags by journalists as they file a story. Subject tags describe the content of the news article and there are over 100 unique subject tags available to journalists such as "US" which is used to label news related to the United States and "DBT" which is used to label debt-related news. For example, an article about US corporate debt would receive both the "US" and "DBT" tags in addition to other tags. For our application, we are interested in isolating all debt-related news articles as identified by the "DBT" subject tag. While the subject tags that accompany news articles allow a forecaster to isolate the type of news that they are interested in, this dataset spans 25 years and may be subject to inconsistent labeling. Chart 1 below shows the average number of subject tags given to news articles at the annual frequency and suggests that subject tagging standards have changed over time. There is a general increase in the number of subject tags given to articles over time with noticeable discontinuity in 2018. Articles published in 2000 have on average eight subject tags while articles from 2018 have on average 39 subject tags.

Chart 1: Average Number of Subject Tags per Article



Note: Average number of subject tags given to news articles by journalists, displayed at the yearly frequency from 1996 to 2021. The average number of subject tags gradually increases over the sample before a noticeable jump in 2018, an indicator of label inconsistency.

The methodology for tagging news articles appears to have changed over time as earlier in the sample journalists had fewer subject tags to choose from. Even if this is not the case, there is still the question of concept drift and whether the relation between text and specific tags has changed over time. In order to address labeling concerns, we propose a method for identifying whether labels are inconsistent over time and backfilling article labels to correct for inconsistent labeling.

Solution to Inconsistent Labeling

We treat the problem of inconsistent labels as a two-step problem. In the first step, a forecaster selects a sample of "ground truth" observations and accompanying labels which reflect consistent labeling standards. If a forecaster is dealing with pre-labeled text, they may select recent observations from the dataset which reflect current labeling standards. If the forecaster is using unlabeled text, they may select a subset of observations and label them manually according to their desired labeling methodology to obtain “ground truth” observations. In the second step, the forecaster fits a classification model on these observations and applies it outside of their training sample to backfill the entire dataset with consistent labels. In this sense, the forecaster’s objective is to create a model which is capable of reading text and producing consistent labels. If the backfilled labels produced by the model differ significantly from pre-existing labels, then labeling inconsistency has been identified and the inconsistent labels may be replaced or augmented with backfilled labels. Our proposed solution depends on traditional structural modeling where the “error” term is formally modeled and used to get to the underlying data generating process. However, in contrast to traditional statistical techniques we utilize pure prediction models to implement our solution. Our proposed solution combines methods from both prediction and attribution and provides an example of how these methods may be practically combined. While our proposed framework allows a forecaster to identify and solve problems of inconsistent labeling, the question of what model should be used to classify news articles in the second step remains. This question is of paramount importance as the performance of a forecaster’s classification model dictates how accurately they can identify labeling inconsistency and how effectively they can address problems of inconsistent labeling. Even if a forecaster selects a robust set of ground truth observations in step one, an inaccurate classification model trained on this data will fail to replicate consistent labeling standards and result in the misidentification of label inconsistency and inaccurate backfill results. For this reason, we pay particular care to model selection when implementing our proposed solution.

Model Selection

In considering what text classification model to use we return to the tradeoffs between attribution technology and pure prediction technology. Within the domain of text analysis there are many attribution-based and pure prediction methods available to us. For our use case, we ultimately consider one attribution-based method, Distributed Multinomial Regression (DMR), and four pure prediction methods, a feedforward neural network, BERT transformer model, random forests, and XGBoost. Before selecting our preferred model and implementing our proposed solution, we consider the tradeoffs between these different technologies and assess the accuracy of each candidate model. We find that model choice has material consequences for classification accuracy and our ability to correct inconsistent labeling.

Contributions

This paper makes three main contributions. First, we examine whether the promise of superior forecasting performance from pure prediction technology is true in the domain of text analysis. Second, in doing so we quantify the tradeoff between attribution and prediction models. Finally, armed with a superior forecasting technique, we propose a method for identifying and solving issues of inconsistent labeling that combines traditional statistical techniques and pure prediction techniques. As a side-effect, we construct a dataset of consistently labeled debt-related news. Given the proliferation of text datasets and datasets with human labelled features, forecasters across many domains will face the labeling issues that we highlight above. While this paper focuses on text analysis, labeling issues are likely to present themselves in human-labeled datasets ranging from email spam to MRI images. While the underlying objects that are being represented in these datasets is vastly different, they all contain a mapping between features and labels that may be analyzed and augmented using our proposed framework. We think our proposed solution to labeling issues has wide ranging applicability and that our findings can inform how forecasters should think about implementing our proposed framework.

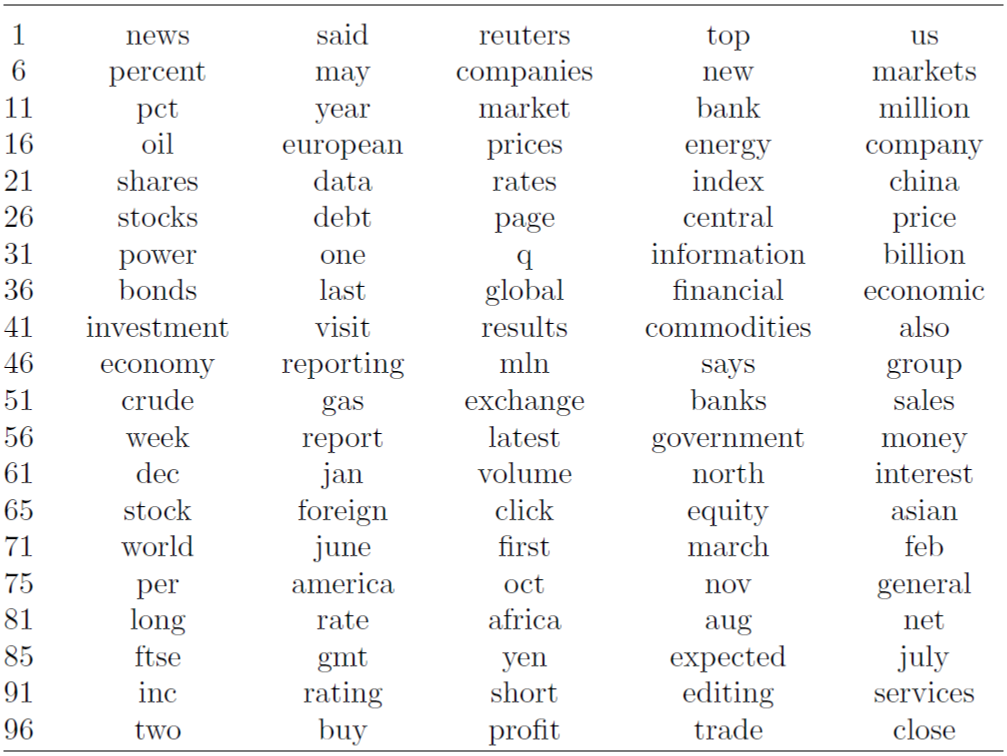
The rest of the paper proceeds as follows. Section 2 describes the data we use, the challenges of using a massive text dataset, and how we formatted our text dataset into a workable format. Section 3 provides an overview of prediction technology and attribution technology, a formal description of our proposed labeling solution, and brief descriptions of the candidate text classification models that we consider. Section 4 describes the framework that we use to evaluate each model and describes performance across our candidate models. Section 5 presents results from our proposed backfill technique and discusses findings. Finally, section 6 concludes.

2. Data

Thomson Reuters Dataset

Our dataset consists of all English language news articles published by the Thomson Reuters news wire service between 1996 and 2021. This dataset is available through the Thomson Reuters News Archive and has been used by researchers such as Londano, Claessens and Correa (2021) who use the dataset to calculate a sentiment index for financial stability communication. We filter the archive for English language news by setting language equal to "EN". We then filter our data for duplicated news articles using the field "last_article" to ensure that we are only picking up the last incidence of each article as some articles are broadcasted multiple times over the window in which the news is deemed relevant. We also note an article type called TOP_NEWS which is a list of news headlines broadcasted multiple times a day. We take the last TOP_NEWS article published each day and remove the other TOP_NEWS articles. After applying these filters, we end up with a corpus of 27 million news articles. Articles in the archive vary in length with some articles containing a simple headline and accompanying sentence and others containing full length news articles with multiple paragraphs. While Thomson Reuters covers many different topics, its audience is primarily the financial community. For context, we display the 100 tokens that appear most frequently in our dataset of all news articles, 41 of which are directly relevant to finance or the macroeconomy. For this reason, we think that the Thomson Reuters dataset is a particularly good source for our objective of creating a debt-related news dataset.

Table 1: Top 100 Most Frequent Tokens 1996 to 2021



1	news	said	reuters	top	us
6	percent	may	companies	new	markets
11	pct	year	market	bank	million
16	oil	european	prices	energy	company
21	shares	data	rates	index	china
26	stocks	debt	page	central	price
31	power	one	q	information	billion
36	bonds	last	global	financial	economic
41	investment	visit	results	commodities	also
46	economy	reporting	mln	says	group
51	crude	gas	exchange	banks	sales
56	week	report	latest	government	money
61	dec	jan	volume	north	interest
65	stock	foreign	click	equity	asian
71	world	june	first	march	feb
75	per	america	oct	nov	general
81	long	rate	africa	aug	net
85	ftse	gmt	yen	expected	july
91	inc	rating	short	editing	services
96	two	buy	profit	trade	close

Note: The 100 tokens (words) that appear most frequently in the Thomson Reuters News Archive between 1996 and 2021. A large portion of the tokens are related to finance or the macroeconomy which reflects the financial focus of the news source.

At the time of publication, Thomson Reuters news articles are labelled with subject tags by the author (a news reporter) that identifies the article's content. We focus on debt-related articles which are labelled with the subject tag "DBT". Articles with the "DBT" subject tag consider both U.S. and non-U.S. debt and are defined by Thomson Reuters as: "Bond and other debt and credit market news, including new issues, yields and other rates, analyses, and market performance. Credit and corporate debt; government, agency, municipal and supranational debt; eurobonds and asset-backed debt." While the subject tags provided by Thomson Reuters provide a method for quickly filtering news by subject, the tagging standards by which news articles are labeled appear to have changed over time. For example, the following article about municipal debt was not labeled as debt-related when it was published in July 2004, however, from the description of the "DBT" tag we would expect it to be identified as debt-related.

> NEW YORK, July 28 (Reuters) - New Jersey's debt was downgraded one notch on Wednesday by a third credit ratings agency, Moody's Investors Service, which expects the state's finances will underperform those of its peers despite its stronger economy. Moody's also noted the state's plan to borrow as much as $1.93 billion in explaining its decision to cut the state's general obligation debt to "Aa3". On Tuesday, Fitch Ratings and Standard & Poor's also cut their ratings on New Jersey's debt by one notch. All three ratings agencies were troubled by Democratic Gov. James McGreevey's plan to close a budget deficit by selling $1.93 billion of bonds. Republican minority leaders had sued to block the sale, arguing it violated the state constitution's balanced budget amendment. The Supreme Court of New Jersey on Monday agreed with the Republicans but decided it would be too disruptive to block the bond sale, a cornerstone of the state's $28 billion budget. But the state's top court did bar governors from ever selling deficit bonds again. Noting New Jersey enacted a $2.9 billion spending hike in the budget that began on July 1, Moody's added: "... The state's budget problems are likely to continue despite the commencement of (the) state economic recovery which began in 2003, and will significantly lag the fiscal recovery seen in other states in the Aa2 category."

Though we cannot confirm label inconsistency from one anecdotal example, the changes in the average number of subject tags we observe in Chart 1 and examples such as the article above raise the question of whether there are additional articles from earlier in the sample that would be considered debt-related by today’s standards but were not tagged as debt-related when they were published. These articles would be missed by a forecaster who simply filtered articles by the “DBT” subject tag and would rob the forecaster of important debt-related information. For this reason, we aim to identify whether labels are inconsistent and address inconsistent labeling.

Text Processing

In order to implement step two of our proposed solution and backfill subject tags, we must first transform our dataset of news articles from raw strings of text into a numeric format that our candidate text classification models may be trained on. Given that we are dealing with over 27 million articles, working with text at this scale is as much a big data problem as it is an estimation problem. The majority of our candidate models take a document term matrix as input as this numeric representation of text is easily machine operable. A document term matrix is a sparse matrix in which *n* rows represent individual news articles and *m* columns contain counts of the number of times that a particular word appears in a given article. This matrix is sparse as authors have many words to choose from and word choice is roughly distributed according to Zipf’s law (see Zipf (1935) and Ryland et. al (2015) for a more recent examination). To produce a document term matrix, we first transform the raw text string of each article into a list of unigram tokens (individual words), remove stop words (commonly used words in English such as “This”, “The”, “At” etc.), and remove special characters. However, we do not stem tokens as this can alter the meaning of a word. For example, we keep “waited”, “waiting”, and “waits” as unique tokens instead of converting them all to “wait”. After tokenizing the entire article, we keep only the 5000 tokens that are used most frequently across all articles published in the 1996 – 2021 period. For a given article, the number of times that each of these 5000 tokens appears is counted and this sparse vector of counts is added as a row in the document term matrix. The final document term matrix of all Thomson Reuters news articles in our dataset has 27 million rows and 5,000 columns, each row representing an article and each column indicating how many times a token appears in a given article. In order to clean and compile text at this scale we utilize Apache Spark, PySpark, and massive parallelism. Simple tasks such as removing stopwords or finding the 5000 most frequently used words become massive computation tasks when performed across 27 million articles. We utilize Apache Spark's functions such as *StopWordsRemover* and *CountVectorizer* which are optimized for distributed data and computation to make these tasks tenable. Data cleaning and compilation is performed across 384 cores and 2,340 gigabytes of memory. Having described our dataset, we next consider in detail the distinction between attribution and pure prediction technologies, formally describe our framework for identifying and addressing inconsistent labeling, and briefly describe the candidate models that we consider for our application.

3. Framework

To clearly distinguish between how our candidate models fit as attribution models or pure prediction models, we describe the distinction between these technologies. We then present our modeling approach and proposed solution to inconsistent labeling. The notation used in our framework closely follows Efron (2020). The data d has n observations of a p-dimensional vector of predictors x_i and associated real valued responses y_i. A general attribution-based model assumes there is some known functional form s such that:

$y_i = s(x_i, \beta) + \varepsilon_i$

Scientists wish to learn this functional form s, which can also be interpreted as x_i causing y_i through the functional form s. As long as n, the number of observations, is larger than p, the number of predictors, there is the possibility of estimating s as there are enough degrees of freedom to estimate a linear functional form. Of course, s could have a non-linear functional form in which case indeterminacy is a problem. A pure prediction model does not consider the existence of s, the underlying function mapping x_i to y_i. “A prediction algorithm is a general program for inputting a dataset d … and outputting a rule $f(x_i, d)$ that, for any predictor vector x, yields a prediction $y^* = f(x_i, d)$.”

$y^*_i = f(x_i, d)$

Models whose focus is on f are pure prediction technologies such as neural networks, random forests, and XGBoost. These models vary in form, but all focus on providing a good prediction, i.e., y* that is close to y. With pure prediction models, there is no pretense that the functional forms that map x_i to y_i, f and s, are similar to each other. In contrast to attribution-based models, the relative size of n and p does not matter. Increases in the number of predictors p generally leads to better prediction and pure prediction models encourage more observations n as the higher the n the more chances the model has at fitting a candidate functional form. This can lead to the problem of overfitting, but having realized this scientists typically withhold part of the sample and use this “out-of-sample” data to test the true reliability of the algorithm. Attribution technologies and pure prediction technologies share a combined hope of mapping x_i to y_i, however, attribution models attempt to find s (the underlying mechanism) while pure prediction models attempt to find f (the most accurate mapping from x_i to y_i).

We follow Athey et al.’s prescription that there is much to be gained from combining pure-prediction technologies with structural modeling techniques developed in the paradigms of traditional statistics. In our case, the problem of inconsistent labeling or concept drift introduces another source of noise to the problem of classifying news articles. As modelers, we get to observe $y^\wedge_i$, human labels, which is y_i mixed with D, the true article labels mixed with distortion D that comes from inconsistent labeling or concept drift. As the modeler cannot observe the true y_i, pure prediction technologies that do not account for D will be hampered as the model optimizes over mapping x_i to a noisy $y^\wedge_i$.

$y^\wedge_i = f(x_i, \beta) + \varepsilon_i + D$

In the worst case, this can devolve into “garbage in garbage out” as pure prediction technology will attempt to minimize the distance between $y^\wedge_i$ and $y*_i$, the human labels and predicted labels. However, using insights from attribution-based technology, we can create a model that does not suffer from distortion D and get most out of pure prediction technology. Based on Chart 1 and associated discussion, we hypothesize that in our dataset distortion D is unknown before 2018 but is zero after that.

Thus, for data after 2018 we assume:

$y_i = y^\wedge_i = f(x_i, \beta) + \varepsilon_I$

Once we have identified a model for the data, s or f, we use our model prediction $y*_i$ to get a list of article labels without distortion D for the entire sample. By comparing our predictions without distortion $y*_i$ to predictions with distortion $y^\wedge_i$ we are able to obtain an estimate of D prior to 2018 and correct for this distortion using our predictions $y*_i$. In this framework, our ability to estimate and correct for D depends crucially on the accuracy of our model, f. Without an accurate model, predictions $y*_i$ will have no relation to the true labels y_i and any comparison with y^ᵢ will produce misleading results. We next turn to a consideration of what model f should be utilized in our particular application.

Candidate Models

As forecasters in the domain of textual analysis, we have many attribution-based and pure prediction models available to choose from when implementing our proposed solution. In the attribution-based camp, some of the most common techniques available to us include custom dictionary approaches (see for example Laver, Benoit and Garry (2003)) or statistical language models (for example Hansen, McMohan, & Pratt, (2018)). However, dictionaries can be brittle as transferring word usage from one domain to another can be confusing (see for example Laughran and McDonald (2011)) and statistical language models depend on subjective choices made by the forecaster such as how many topics a document has. An important breakthrough in attribution technology came with Taddy (2013) which builds upon Cook's (2007) work on inverse regressions and introduces sufficient reduction for textual data. Taddy (2015) further extends this work by providing a way to compute these statistics in parallel via Distributed Multinomial Regression (DMR). The DMR model and the sufficient reduction statistic that it generates captures all of the information in text for a target variable, in this case the label of the article. This provides comfort as a forecaster is guaranteed to have captured all relevant information in the text given some mild assumptions. We consider DMR as the state-of-the art benchmark for traditional statistics-based attribution technology, and include it in our list of candidate models. While attribution models such as DMR have a number of attractive features, the same labeling task can be accomplished with a myriad of pure prediction technologies such as neural networks or random forests. Outside of the paradigms of traditional statistics these models are the go-to option for classification problems, and we would be remiss to not consider them for our proposed solution. We include four pure-prediction models, a feedforward neural network, BERT transformer model, random forests, and XGBoost to our list of candidate models. Table 2 below shows the models we evaluate in our classification problem and identifies them as attribution-based models or pure prediction models.

Table 2: Candidate Models

	DMR	Feedforward	BERT Model	Random Forest	XGBoost
Attribution Based	X				
Pure Prediction		X	X	X	X

Note: Classification of each candidate model as an attribution-based model or pure prediction model.

One important question is why we do not consider a simple linear regression, the most traditional attribution-based model, in our list of candidate models. A linear regression model is not identifiable in our case as the number is predictors is too large and is thus omitted. We briefly describe each of our candidate models next.

DMR Model

Given that the DMR model is our only attribution-based model, we describe it in greater detail than the other models. The Distributed Multinomial Regression (DMR) model (Taddy 2015) develops on his earlier work on multinominal inverse regressions (Taddy 2013). This model takes a sparse document term matrix of *n* rows and *m* columns, one row for each news article and one column for each token. Rows of the matrix are filled with counts representing the number of times each token appears in a given article. The DMR model also takes a covariate matrix of variables related to token counts. In our case, this is a binary vector representing whether an article has the ‘DBT’ subject tag or not. For each token (column), gamma lasso regularization is used to fit an inverse Poisson regression of the token count on the ‘DBT’ tag covariate.

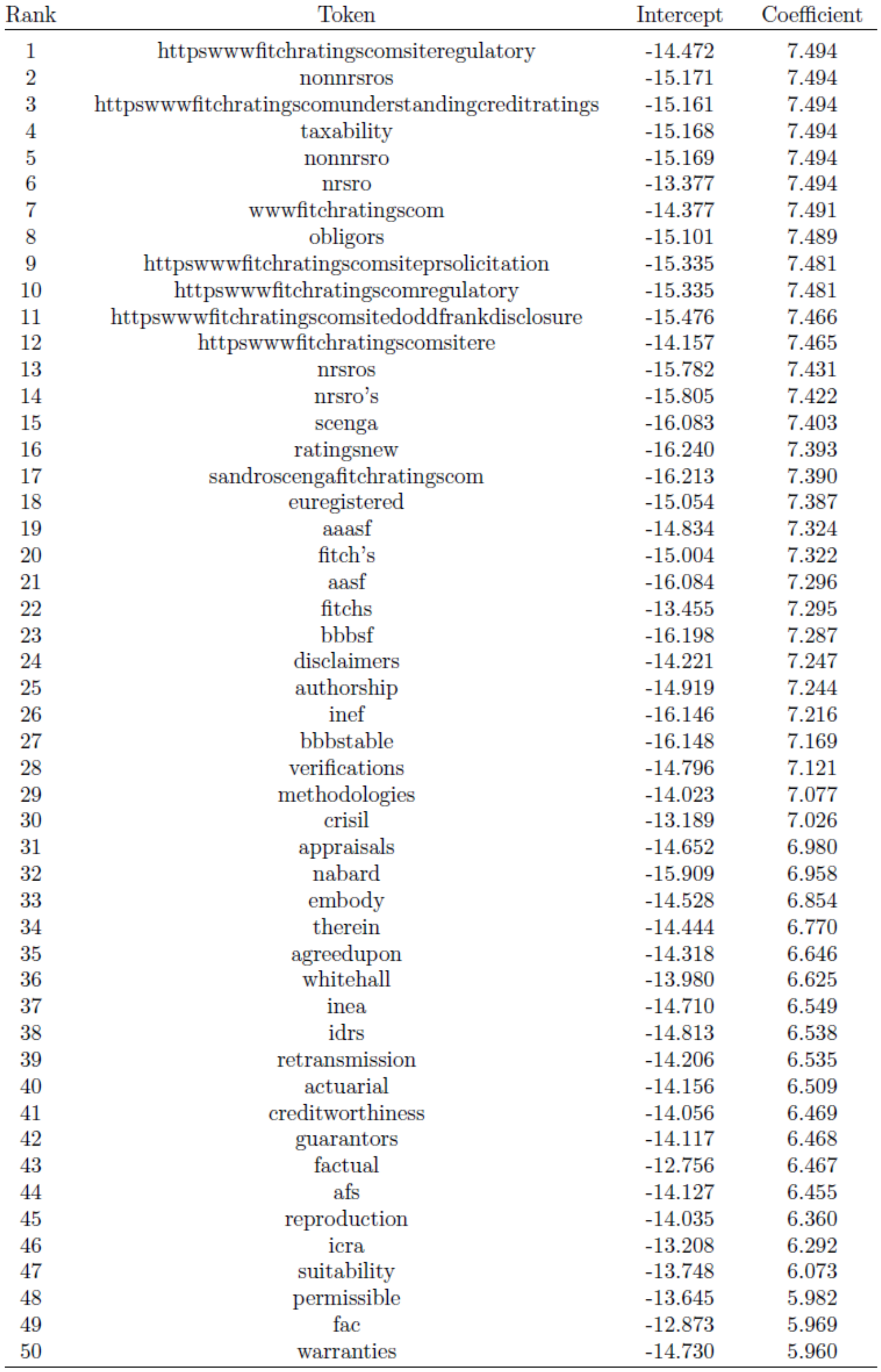
Token ~ DBT

Coefficients from each inverse Poisson regression are then multiplied by the token counts of a given article and scaled by the sum token count of the article. This yields a single sufficient reduction statistic for each article that is based on the token counts of the article. Sufficient reduction statistics (Cook, 2007) for multinomial predictors have the feature that y_i is independent of x_i given the sufficient reduction. This leads to massive dimension reduction from a vector of all tokens in a news article to one sufficient reduction number. Given some assumptions about the distribution of token counts, this sufficient reduction number captures all information in the article with respect to the covariate. After computing the sufficient reduction statistic of each article, we use a logit regression of the 'DBT' tag on the sufficient reduction statistic to classify articles as debt-related or non-debt-related. Unlike our other models, the DMR model is interpretable. One can observe token coefficients and see which tokens are most or least likely to result in an article being classified as debt-related or non-debt-related. The ability to view coefficient magnitudes and reduce text to a single sufficient reduction statistic places the DMR model firmly in the camp of traditional attribution techniques. We use the R package distrom to implement the DMR model.

Insights from DMR Estimation

Our lone attribution model permits us a unique view into what information is used to classify news as debt-related. Table 3 below shows the top 50 (by coefficient magnitude) tokens from the DMR model estimation.

Table 3: Top 50 DMR Coefficients by Magnitude



Rank	Token	Intercept	Coefficient
1	httpswwwfitchratingscomsiteregulatory	-14.472	7.494
2	nonnrsros	-15.171	7.494
3	httpswwwfitchratingscomunderstandingcreditratings	-15.161	7.494
4	taxability	-15.168	7.494
5	nonnrsro	-15.169	7.494
6	nrsro	-13.377	7.494
7	wwwfitchratingscom	-14.377	7.491
8	obligors	-15.101	7.489
9	httpswwwfitchratingscomsiteprsolicitation	-15.335	7.481
10	httpswwwfitchratingscomregulatory	-15.335	7.481
11	httpswwwfitchratingscomsitedoddfrankdisclosure	-15.476	7.466
12	httpswwwfitchratingscomsitere	-14.157	7.465
13	nrsros	-15.782	7.431
14	nrsro's	-15.805	7.422
15	scenga	-16.083	7.403
16	ratingsnew	-16.240	7.393
17	sandroscengafitchratingscom	-16.213	7.390
18	euregistered	-15.054	7.387
19	aaasf	-14.834	7.324
20	fitch's	-15.004	7.322
21	aasf	-16.084	7.296
22	fitchs	-13.455	7.295
23	bbbsf	-16.198	7.287
24	disclaimers	-14.221	7.247
25	authorship	-14.919	7.244
26	inef	-16.146	7.216
27	bbbstable	-16.148	7.169
28	verifications	-14.796	7.121
29	methodologies	-14.023	7.077
30	crisil	-13.189	7.026
31	appraisals	-14.652	6.980
32	nabard	-15.909	6.958
33	embody	-14.528	6.854
34	therein	-14.444	6.770
35	agreedupon	-14.318	6.646
36	whitehall	-13.980	6.625
37	inea	-14.710	6.549
38	idrs	-14.813	6.538
39	retransmission	-14.206	6.535
40	actuarial	-14.156	6.509
41	creditworthiness	-14.056	6.469
42	guarantors	-14.117	6.468
43	factual	-12.756	6.467
44	afs	-14.127	6.455
45	reproduction	-14.035	6.360
46	icra	-13.208	6.292
47	suitability	-13.748	6.073
48	permissible	-13.645	5.982
49	fac	-12.873	5.969
50	warranties	-14.730	5.960

Note: Top 50 token coefficients estimated by the DMR model. The presence of these tokens in a news article is likely to result in the news article being classified as debt-related. These tokens indicate that the DMR model is using sensible features of the text data to classify articles as debt-related or non-debt-related.

These are the intercepts and coefficients estimated in the inverse Poisson regression, token ~ DBT, that is fit for each token. The coefficient indicates how many times a token is likely to appear in an article given that the article is labeled as debt-related or non-debt-related, with larger coefficients and intercepts corresponding to tokens that are more likely to appear in debt-related articles. In a given article individual tokens appears with a very low probability due to the sheer number of words in the English language, something that is captured by the negative constant terms in the inverse Poisson regressions. Take for example the token "taxability" which has a coefficient of 7.494 and an intercept of -15.168. In articles labeled as debt-related, "taxability" is estimated to appear 0.000464 times.[4] While this is a low number, it is much higher than the 0.000000258 times that "taxability" is expected to appear in articles that are not debt-related. The ratio of expected counts for debt-related articles to non-debt-related articles is e^(coefficient). For example, the word "taxability" is e^7.49 or 1,797 times more likely to appear in an article if it is debt-related than if it is not debt-related. With this context, we consider the tokens that the DMR deems most important for classifying news articles as debt-related. The token with the largest estimated coefficient is "httpswwwfitchratingscomsiteregulatory". This Fitch domain name is sensible as Fitch is a large rating agency, and unlike its competition S&P it predominately covers debt. Tokens #1, 3, 7, 9, 10, 11, 12, 20, and 22 all feature "Fitch", indicating that the existence of Fitch related tokens in an article is a good indicator that the article is debt-related. Tokens #2, 5, 6, 13, and 14 also pertain to a rating agency, but refer to "nrsro" which stands for "Nationally Recognized Statistical Ratings Organization". The first prominent token not associated with a rating agency is token #4, "taxability", which is a

[4] The expected number of occurrences of a token in an article may be computed as: e^(intercept + coefficient * DBT). For the token "taxability", given that an article is labeled as debt-related with the DBT tag, the expected number of occurrences of the token "taxability" in the article is: e^(-15.168 + 7.493 * 1) = 0.000464.

prominent concern for debt instruments. The second such token is token #8, “obligators”, which is often how firms or entities that take on debt are referred to. Another prominent token is token #15, “scenga”, which is the last name of Fitch’s media relations person – Sandro Scenga. His name appears again in token #17, this time with his full name and email address sans special characters which are removed in our text cleaning process. Other rating agencies also appear in the top coefficients, for example “crisil” at token #30, and “icra” at token #46. Accounting related terms start to appear further down the list such as “verifications” at token #28, “actuarial” at token #40, and “credit worthiness” at token #41. The coefficients we consider above give a concrete example of some of the benefits and limits of attribution models in the domain of text analysis. Attribution models provide a forecaster the comfort of verifying that parameters estimated by a model are sane. They also permit the forecaster to discover potentially interesting relationships such as the link between “taxability” and debt-related news identified above. However, because of the large number of predictors in textual data, many relationships are uncovered that may not be interesting to a forecaster such as the connection between Fitch website links and debt-related news. The parameters that lead to good prediction may not lead to compelling attribution. Furthermore, it is unlikely that a forecaster will be able to interpret all predictors. In the example above we only look at the intercept and constant of 50 of the 5000 tokens in our document term matrix, and it is unrealistic for a forecaster to consider all 10,000 parameters. We next describe the four pure prediction models considered in this study. Given that these models are complex and that their architecture is not the focus of this paper, descriptions are brief. For each model, the exact hyperparameters we use when implementing the model may be found in the appendix.

Feedforward Neural Network

A classic feedforward neural network uses a large number of weights and biases, non-linear activation functions, gradient descent, and backpropagation to produce predictions for a target variable. These features make neural networks a prime example of pure prediction technology. We train our neural network with the same document term matrix described above using the following neural network architecture: 5000 neuron input layer, ReLU activation, 100 neuron hidden layer, ReLU activation, 10 layer output layer, sigmoid activation. We train the model using a mean squared error (MSE) loss function and stochastic gradient descent with a learning rate of 0.01. We use the Python module *PyTorch* to train our neural network.

BERT Model

BERT is a transformer model which takes tokenized article text as input instead of a document term matrix. Unlike other pure prediction technologies such as neural networks and random forests that can be applied across domains, BERT is designed with textual data in mind. Whereas our other models only see the number of times a token appears in an article, the BERT model sees where each token appears in the article and the context in which the token appears. The BERT model is also unique in that it has memory of past inputs. The cost of these features is magnitudes greater computation time. This constraint led us to handicap the BERT model in a number of ways so that computation time would be tenable.[5] For greater detail on BERT see Devlin et al. (2018). We use the Python module *transformers* and the pre-trained *DistilBert* model.

[5] Constraints include using the DistilBERT model instead of the full BERT model and restricting the model to only look at the first 100 tokens of a given article instead of the entire article.

Random Forests

Random forests is a decision-tree based ensemble method. Our document term matrix is bootstrapped and then used to create multiple decision trees based on token counts. Each decision tree uses a stochastically selected subset of the tokens to classify articles as debt-related or not debt-related. After articles have been classified by each decision tree, the final class prediction is the class predicted by the majority of decision trees. Given that the classification thresholds of each decision tree are observable one may argue that random forests fall in the category of attribution technology. However, given the sheer number of decision trees and the fact that features are stochastically selected, attribution is unrealistic. It is the wisdom of many trees that gives random forests added value over an individual decision tree, a clear indicator that random forests is a pure prediction model. For greater detail on random forests see Breiman (2001). We use the *ranger* R package to implement our random forests model.

XGBoost

The last model we consider is another tree-based ensemble method, XGBoost. Similar to random forests, XGBoost uses decision trees to classify articles based on the counts of individual words in our document term matrix. However, the XGBoost model implements additional trees to predict the residuals of prior trees and uses a gradient descent algorithm to minimize a loss function when adding decision trees. The final prediction of the XGBoost model is a weighted sum of all tree predictions. The usage of gradient descent puts XGBoost models firmly in the camp of pure prediction models. XGBoost models are explained in detail in Chen and Guestrin (2016). We use the *xgboost* R package to train our XGBoost model.

4. Model Evaluation

Having introduced our candidate models, we turn to evaluating which model is best for our proposed backfill solution. Although attribution-based models and pure prediction models have radically different objectives, both classes of model can be evaluated head-to-head based on their prediction ability. While we have noted the benefits of interpretation and comfort that attribution models provide, for our use case we are interested in the model that best emulates journalist tagging standards in the later years of the Thomson Reuters dataset. To assess this, we train each of our text classification models on all news articles published in 2018. We then test these models out-of-sample on all articles published in 2017 and 2019 as well as in-sample on all articles published in 2018 to examine how well each model replicates journalist tags. Recall that in order to implement our backfill solution, we must train a text classification model which accurately replicates human tagging standards during a period with consistent tagging standards. An effective model should tag debt-related and non-debt-related articles similarly to how journalists tagged them. Articles published in 2018 are on average labeled with 39.78 subject tags. This gives us confidence that articles from 2018 are reflective of current tagging standards and that this is an appropriate sample to train our models on. Articles published in 2019 also appear to be robustly tagged with an average of 40.88 subject tags. Articles published in 2017, however, appear to be less thoroughly tagged with only 16.18 subject tags on average. By testing our model out-of-sample on articles from 2017 and 2019 separately, we get a sense of whether the performance of each model changes depending on the time period in which it is tested. It is important for our backfill model to perform well in earlier periods that are sparsely tagged, and testing out-of-sample on articles from 2017 allows us to assess this.

Model Accuracy Metrics

In our model test, a true positive occurs when a model predicts that an article is debt-related and it has indeed been labeled as debt-related by the author of the article. A false positive occurs when a model identifies an article as debt-related but it was not identified as debt-related by the author of the article. Classifying debt-related articles is an imbalanced classification problem as only 11% of all news articles are labeled as debt-related. This presents evaluation issues as a naïve model which classifies all articles as non-debt-related would achieve a headline accuracy of 89%, a reasonable success rate but a useless model. Due to this imbalance, we evaluate model performance based on precision, recall, and F1 score. These measures are defined as follows.

$$\text{Precision} = \text{True Positive} / \text{True Positive} + \text{False Positive}$$

$$\text{Recall} = \text{True Positive} / \text{True Positive} + \text{False Negative}$$

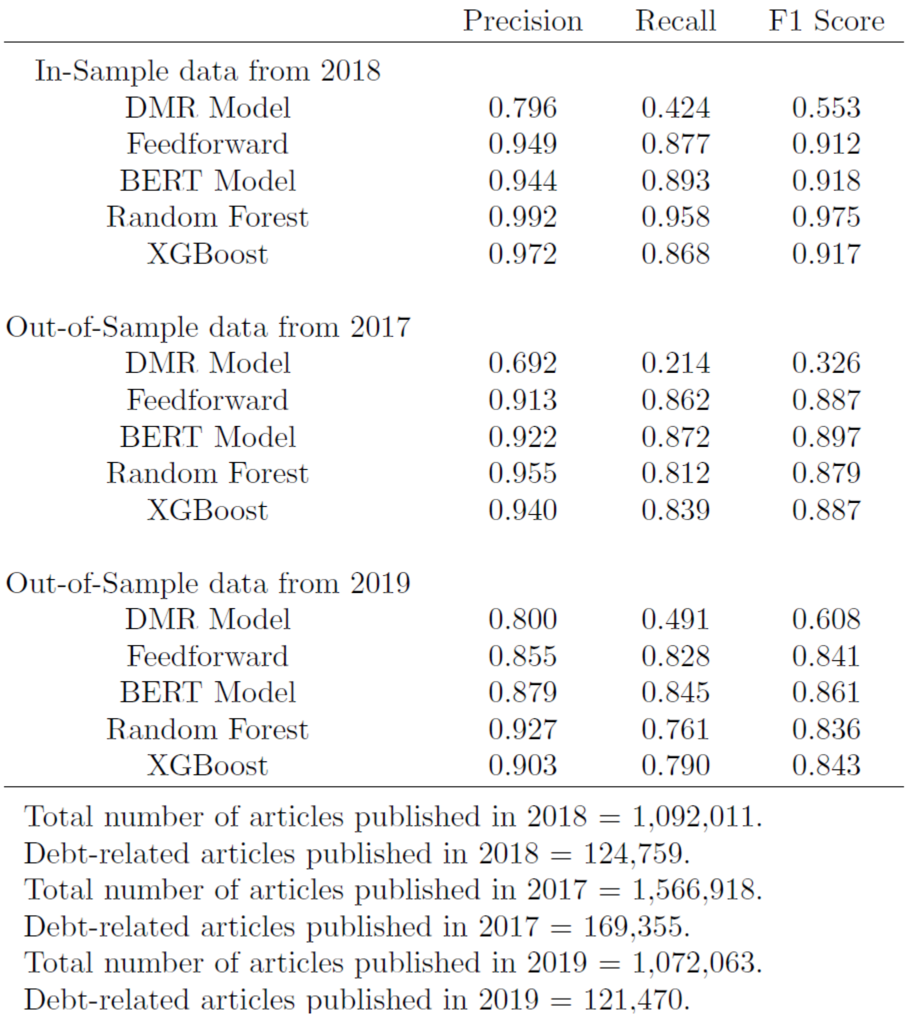
$$\text{F1 Score} = 2 * \text{Precision} * \text{Recall} / \text{Precision} + \text{Recall}$$

Our objective is to identify as many debt-related articles as possible without falsely identifying extraneous articles, and our choice of evaluation statistics reflects this joint optimization problem. We consider three widely used measure – precision, recall and F1 score. Precision, the ratio of true positives to true positives and false positives, penalizes models for identifying extraneous articles. Precision asks the question, "of all articles that the model predicted as debt-related, what percentage were actually debt-related?" Recall, the ratio of true positives to true positives and false negatives, penalizes models for not identifying debt-related articles. Recall asks the question, "of all articles that are actually debt-related, what percentage were identified by the model?" F1 score, the harmonic mean of precision and recall, combines these measures. For our application, we value Recall as the most important measure as it judges whether a model will identify debt-related articles when it encounters them, even if high Recall comes at the cost of including a few additional false positives. For all three evaluation statistics, higher scores represent better model performance.

Performance Across Models

Results of our model comparison are contained below in Table 4.

Table 4: Classification Accuracy



	Precision	Recall	F1 Score
In-Sample data from 2018			
DMR Model	0.796	0.424	0.553
Feedforward	0.949	0.877	0.912
BERT Model	0.944	0.893	0.918
Random Forest	0.992	0.958	0.975
XGBoost	0.972	0.868	0.917
Out-of-Sample data from 2017			
DMR Model	0.692	0.214	0.326
Feedforward	0.913	0.862	0.887
BERT Model	0.922	0.872	0.897
Random Forest	0.955	0.812	0.879
XGBoost	0.940	0.839	0.887
Out-of-Sample data from 2019			
DMR Model	0.800	0.491	0.608
Feedforward	0.855	0.828	0.841
BERT Model	0.879	0.845	0.861
Random Forest	0.927	0.761	0.836
XGBoost	0.903	0.790	0.843

Total number of articles published in 2018 = 1,092,011.
Debt-related articles published in 2018 = 124,759.
Total number of articles published in 2017 = 1,566,918.
Debt-related articles published in 2017 = 169,355.
Total number of articles published in 2019 = 1,072,063.
Debt-related articles published in 2019 = 121,470.

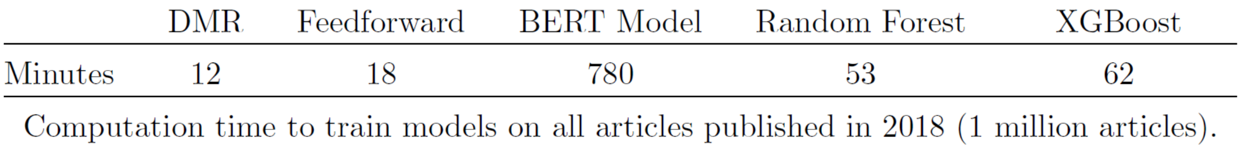
Note: Precision, recall, and F1 scores of our candidate models evaluated in-sample and out-of-sample on data from 2017, 2018, and 2019. Across samples and metrics, the attribution-based DMR model is outperformed by our four pure prediction models with the feedforward neural network and BERT model performing best over our favored criterion, recall.

The DMR model returns precision scores that are only slightly lower than the other four models as it excels at not predicting false positives, articles that are labeled as debt-related by the model but not by the journalist.[6] Testing on 2017 (2019) data, the DMR model predicted 16,066 (14,944) false positives, the Feedforward neural network predicted 13,979 (17,105) false positives, the BERT model predicted 12,469 (14,179) false positives, the random forests model predicted 6,443 (7,276) false positives, and the XGBoost model predicted 9,033 (10,290) false positives. In terms of precision, the DMR model beats the feedforward neural network in 2019 and is on par with the BERT model. The XGBoost and random forests models outperform all other models at not identifying false positives, with the random forests model being the clear winner. While the precision of the DMR model is worse than many of the pure prediction models, the number of false positives that it identifies is within a similar magnitude. Given that over a million news articles were published both in 2017 and 2019, a forecaster may be willing to falsely identify 8000 additional debt-related articles for the benefits of comfort and attribution that the DMR model provides. However, the precision of the DMR model comes at the cost of recall. In 2017 (2019) the DMR model predicted 36,167 (59,620) true positives, the feedforward neural network predicted 146,452 (100,626) true positives, the BERT model predicted 148,266 (102,634) true positives, the random forests model predicted 137,881 (92,517) true positives, and the XGBoost model predicted 142,680 (96,011) true positives. The DMR model identifies significantly fewer debt-related articles than all of the pure prediction models in both 2017 and 2019. Additionally, given our objective of backfilling sparsely tagged articles from earlier in our

[6] Model performance is conditional on the hyperparameters chosen by the forecaster such as learning rate, number of epochs, number of trees, etc. When possible, we opt for the default hyperparameters specified by the software packages that we use to implement these models. However, we find that across models results are robust to changes in hyperparameters. The exact hyperparameters we use may be found in the appendix.

sample, it is worrisome that the DMR model performs significantly worse out-of-sample in 2017 than in 2019. In contrast, the four pure prediction models perform better in 2017 than in 2019. Considering that we weight recall as the most important metric of model performance, the low recall scores in the DMR rule it out as a possible model for implementing our proposed labeling solution. In comparing the recall of our pure prediction models, the BERT model and feedforward neural network outperform the random forests and XGBoost models out-of-sample, with the BERT model having the best recall by a slight margin. Given that we put the most weight on the ability of a model to identify debt-related articles, the results of these tests leave the feedforward neural network and BERT model as our two favored models. Before backfilling subject tags, the additional criterion of compute time must be considered. Table 5 below shows the number of minutes it takes to train each model on all articles published in 2018, roughly 1 million news articles.[7]

Table 5: Computation Costs



	DMR	Feedforward	BERT Model	Random Forest	XGBoost
Minutes	12	18	780	53	62

Computation time to train models on all articles published in 2018 (1 million articles).

Note: Time in minutes that it takes to train each of our candidate models on all news articles published in 2018, roughly 1 million articles. All models are relatively inexpensive to train in comparison to the BERT model which takes a significant amount of time to train.

The DMR model and feedforward neural network are both cheap to train, the random forest and XGBoost models are more expensive, and the BERT model is magnitudes more expensive than any of the other models to train. Considering compute constraints and model performance in the

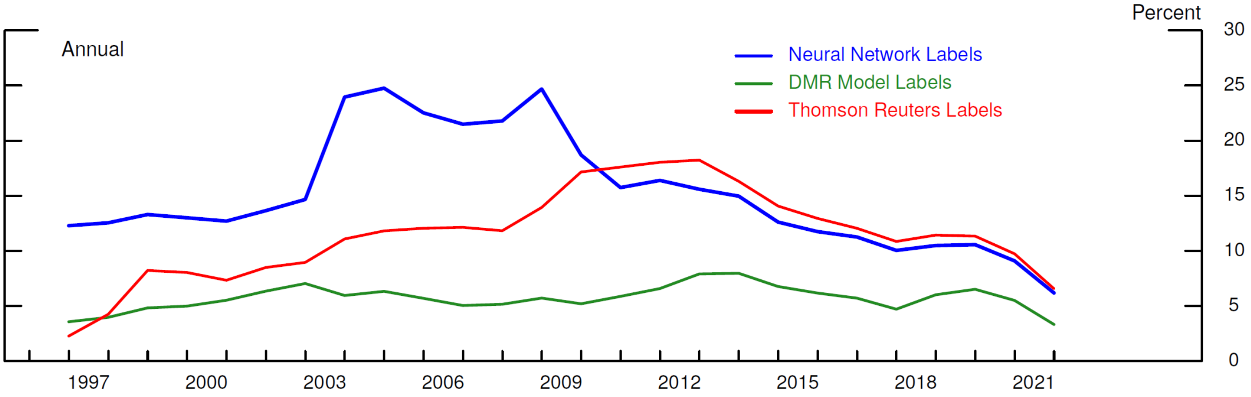
[7] Just as with model performance, compute time is conditional on the model hyperparameters chosen by the forecaster.

2017 to 2019 period, we select the feedforward neural network as our preferred model for backfilling subject tags. Although we ultimately elect for the feedforward neural network as our model of choice, random forests or XGBoost would have also been good choices for this task.

5. Estimates of Label Inconsistency

Having selected our preferred model, we turn to implementing our proposed solution. In this step, we train our preferred feedforward neural network on all news articles from 2018 to 2021 and subsequently use this model to identify debt-related articles over the entire 1996 to 2021 period. We also implement the same steps using the DMR model in order to quantify the difference between a pure-prediction-based approach and an attribution-based approach. Articles in our 2018 to 2021 training sample have on average 40.72 subject tags and reflect current labeling standards, thus we assume that inconsistent labeling distortion D is zero in this period. As expected, we find that our neural network identifies significantly more debt-related articles than Thomson Reuters journalists with increases in the number of debt-related articles concentrated in earlier periods when we suspect that the subject tagging methodology used by journalists was most different from present day standards. It is also likely that concept drift is most prevalent in earlier periods as changes in the mapping between text and subject tags occurs over time, contributing to the discovery of debt-related articles earlier in history. To control for changes in the number of articles published by Thomson Reuters each year, we consider the number of articles labeled as debt-related in a given year divided by the total number of articles published that year. Chart 2 below plots the proportion of all Thomson Reuters news articles that are labeled as debt-related by our neural network (blue), DMR model (green), and Thomson Reuters journalists (red) over time.

Chart 2: Proportion of Articles Labeled as Debt-Related

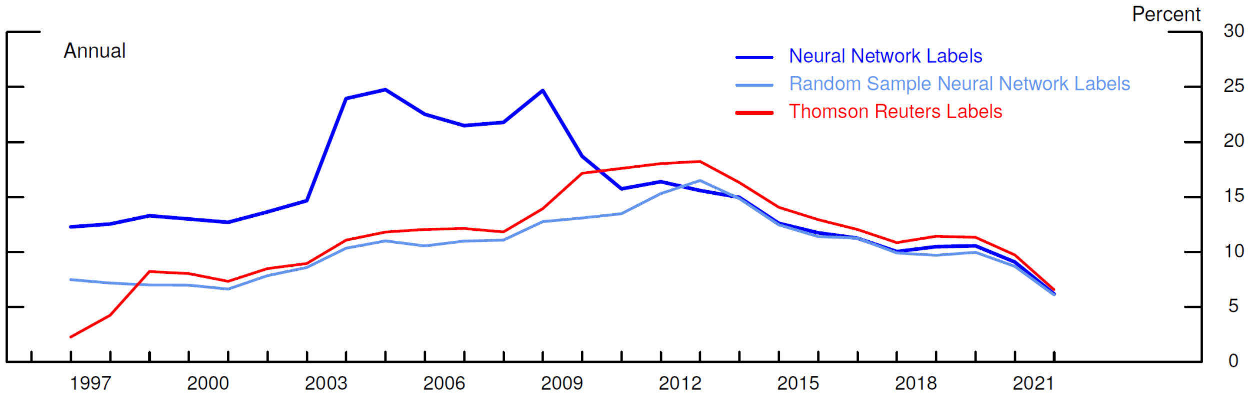


Note: Proportion of all articles published in a year that are labeled as debt-related by our neural network, DMR model, and Thomson Reuters journalists. Differences between neural network labels (blue) and Thomson Reuters journalist labels (red) prior to 2009 indicate that labeling standards were different earlier in the sample before coming into line with current standards later in the sample.

Our neural network identifies significantly more debt-related articles early in the sample when differences in tagging methodology and concept drift are most pronounced, but roughly the same number of debt-related articles later in the sample as labeling standards come into line with what is in our training sample and D decreases. In the 1996 to 2009 period the average proportion of debt-related articles identified by Thomson Reuters journalists is 9.8% (1,256,901 articles in total), while the proportion of debt-related articles identified by our neural network is 17.9% (2,286,108 articles in total), almost double the number of debt-related articles. In the 2010 to 2021 period the average proportion of articles identified as debt-related by journalists and our neural network are similar with journalists identifying 13.3% (1,819,583 articles in total) of articles as debt-related and our neural network identifying 12% (1,657,850 articles) of articles as debt-related. A notable feature of the chart above is the uptick in the number of debt-related that our neural network identifies between 2002 and the Global Financial Crisis in 2008. Though we can only speculate, this uptick may be a result of the fact that prior to 2008 articles mentioning housing prices, subprime mortgages, and other subjects that turned out to be relevant to debt and the economy weren't considered debt-related by journalists. After the Global Financial Crisis labeling standards may have changed to consider these articles as debt-related, causing our neural network to pick up these additional articles. Another notable feature is how few articles the DMR model labels as debt-related. Over the entire 1996 to 2021 sample the average proportion of articles labeled as debt-related by the DMR model is only 5.7%. This is consistent with our finding that the DMR model has low recall, and highlights the fact that our proposed solution to inconsistent labeling is only effective if it is implemented with an accurate model. A researcher using the DMR model for this particular problem would fail to identify many articles considered debt-related by todays labeling standards. Our findings above provide evidence that labeling is inconsistent over time and that our neural network is able to address this issue. If the subject tagging methodology of journalists had not changed over time, we would expect the blue line to remain close to the red line over the entire sample as the neural network would simply identify articles that had already been labeled as debt-related by authors. If the neural network was not effective at replicating subject tagging methodology, we would expect to see a difference between the blue line and red line in the later period as the neural network failed to match to tagging habits of recent article authors, a finding we observe with the DMR model. Another sign of efficacy is that in addition to identifying new debt-related articles, our neural network is effective at identifying the same debt-related articles that were labeled by authors. Over the entire sample, 74% of articles that were identified as debt-related by journalists were also identified as debt-related by our neural network. In contrast, only 32% of articles labeled as debt-related by journalists were identified as debt-related by the DMR model, a result that is complementary to the model recall scores in Table 4. These results hold over the 1996 to 2017 period in which articles have significantly fewer subject tags, with our neural network identifying 72% of author labeled debt articles and the DMR model only identifying 31%. The fact that our neural network identifies the majority of articles labeled as debt-related by journalists implies that articles considered as debt-related by earlier tagging standards are likely to still be considered as debt-related by today's standards. This also implies that the new debt-related articles identified by our neural network are not the result of a fundamental change in what news is considered debt-related, but rather an expansion in the type of news that is considered debt-related. This is commensurate with our hypothesis that the labeling of debt-related news has changed over time as the scope of what news is considered debt-related has increased.

As an additional check of our results, we perform our backfill technique using a neural network that is trained on a random sample of articles from the entire 1996 to 2021 period. Instead of training our neural network on the 4.5 million articles that were published between 2018 and 2021 and reflect present tagging standards, we train our neural network on 4.5 million randomly sampled articles which reflect tagging standards from the whole sample and contain distortion D. The results of this exercise are shown in Chart 3 below with the light blue line representing the proportion of all articles labeled as debt-related by the random sample neural network.

Chart 3: Random Sample Model Labels



Note: Proportion of all articles published in a year that are labeled as debt-related by our neural network, Thomson Reuters journalists, and a neural network trained on a random sample of news articles from the entire 1996 to 2021 sample. The similarity between the random sample neural network (light blue) and Thomson Reuters journalist labels (red) highlights the importance of selecting a training sample with desired labeling standards.

In contrast to our standard neural network (blue) which identifies inconsistent labeling D in earlier periods, the random sample neural network does not identify any inconsistent labeling D with the random sample neural network (light blue) closely tracking journalist labels (red). These findings highlight that an integral part of our proposed solution is identifying a training sample that reflects desired labeling standards, i.e. a training sample where D, the distortion, is as close to zero as possible. The results above also demonstrate the strong recall capability of the neural network. The random sample neural network does a good job of replicating the labeling behavior of journalists with 77% of all articles labeled as debt-related by journalists also being identified as debt-related by the random sample neural network. This is reflected in the similar movement of the light blue and red lines above and serves as further evidence that the neural network is a good model for our proposed solution.

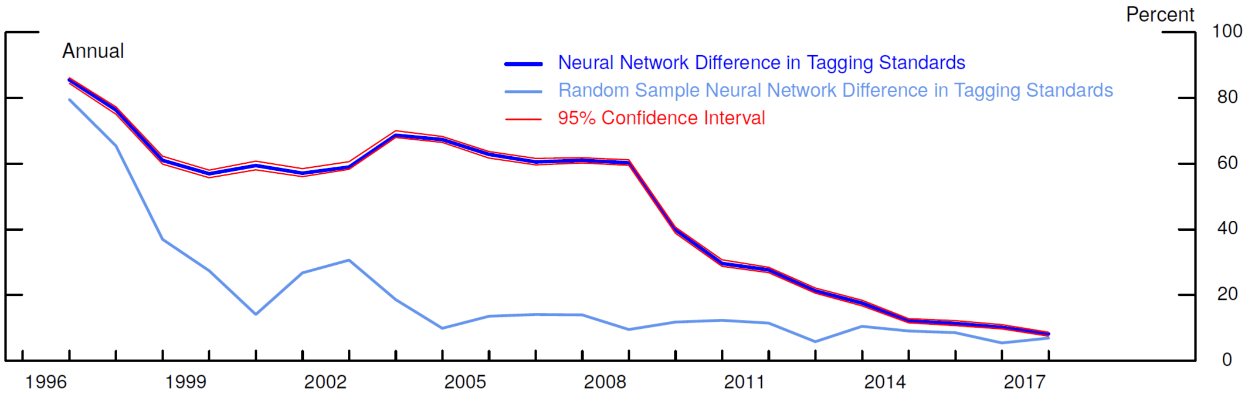
An Estimate of Distortion

From our results we infer that subject tagging has changed over time, and with our proposed backfill technique we have the tools to quantify exactly how labeling has changed. Recall that when motivating the case of Thomson Reuters news articles, we did not have definitive evidence that tagging standards had changed over time or that labels suffered from concept drift. We had evidence there were changes over time in the average number of subject tags per article, however, without hand-reading millions of articles there was no way to verify that subject tagging methodology had changed over time. Our backfill technique provides a solution to this problem and a method for estimating distortion D. If a model is perfectly fit to the tagging standards of a specific period (in the sense that it perfectly replicates the way that a human author tags an article), then one can apply this model outside of the training period and know the degree to which tagging standards have changed over time. While we do not have a model that perfectly replicates human tagging behavior, we know how well our models performs in-sample and we know how many additional articles it identifies out-of-sample. Our neural network trained on articles from the 2018 to 2021 period has an in-sample precision of 0.93, recall of 0.87, and F1 score of 0.90. This provides confidence that our model accurately replicates the subject tagging norms of the 2018 to 2021 period, an assertion that is necessary to measure how tagging norms have changed over time.[8] We measure difference in tagging standards (distortion D) as the number of additional debt-related articles identified by our backfill model divided by the total number of debt-related articles identified by our backfill model. Additional debt-related articles are articles that are labeled as debt-related by our neural network but not labeled as debt-

[8] Note that in-sample F1 score is the most important metric for deriving confidence as it provides a balanced assessment of a model's ability to replicate current labeling standards. A naïve model that classifies all articles as debt-related would achieve a recall score of 1, but would be useless for identifying changes in tagging standards over time.

related by Thomson Reuters authors. If subject tagging has not changed over time, then our neural network will not identify any additional articles and this statistic will equal 0. If subject tagging is radically different, the number of additional articles identified by our neural network will equal to the total number of articles that it identifies, and this statistic will equal 1. In this latter scenario, the overlap between debt-related articles identified by the neural network and journalists is 0. We find that out-of-sample (1996 to 2017) our neural network identifies 3.5 million articles in total, of which 1.6 million are additional debt-related articles. This implies that subject tagging between the 1996 to 2017 period and 2018 to 2021 period changed by 0.46 or 46% on average. We calculate this measure of change at the yearly frequency to see how subject tagging has changed over time. Chart 4 below shows how differences in tagging standards evolve over the 1996 – 2017 period as expressed in percentage.

Chart 4: Difference in Tagging Standards



Note: Difference in tagging standards over time implied by our neural network and random sample neural network. Measured as the number debt-related articles that are identified by each model but not by Thomson Reuters journalists divided by the total number of articles identified as debt-related by the model. Our neural network (blue) indicates that tagging standards were different prior to 2008 before gradually coming into line with current labeling standards later in the sample.

Early in the sample tagging standards are almost entirely different to present-day tagging standards. In 1996, 86% of articles identified by our neural network as debt-related articles were not identified by journalists using the tagging standard of the time. Over the course of our sample the difference in tagging standards decreases. In 2012 tagging standards are 21% different and by 2016 tagging standards are only 10% different. After 2008 the tagging standards used by journalists quickly come into line with the tagging standards used by present-day authors. To check the robustness of these estimates we perform the same exercise with 100 bootstrapped samples of 1 million articles each and construct a 95% confidence interval with the results. We find that our estimates of changes in tagging standards over time are robust with our bootstrapped estimates (red) displaying roughly the same contour as the results from our standard neural network (blue). We also perform this exercise using a random sample neural network (light blue) and find minimal difference in tagging standards apart from very early in the sample. This is expected as the random sample neural network assumes there is no distortion and should mimic journalist labeling. By fitting a model that accurately captures current tagging standards, a forecaster may simultaneously solve the problem of inconsistent labels and demonstrate that the problem existed in the first place. Additionally, a forecaster can see exactly when and to what degree labeling changed over time.

Implications

Having identified and backfilled inconsistent article labels, we find the implications of this task to be significant. After backfilling subject tags with the feedforward neural network, we identify 1.7 million (1,660,963) additional debt-related articles that were not present in the 3.1 million (3,076,484) articles tagged as debt-related by article authors. With our end goal of creating a complete dataset of debt-related news in mind, we ultimately take the union of articles identified as debt-related by Thomson Reuters journalists and our neural network as our final debt-related news dataset. This results in a dataset with 4.7 million news articles (4,737,447), a 54% increase in dataset size relative to the debt-related news dataset that would have been acquired by a researcher who ignored inconsistent labeling issues and simply filtered by the pre-existing "DBT" label. Even if we had decided to only take articles identified as debt-related by our neural network when constructing our final dataset, we would still end up with a dataset of 3.9 million (3,943,958) news articles, a 28% increase in observations relative to a dataset that takes journalist labels as given. Our results also emphasize the importance of model selection. Had we relied on attribution-based methods (the DMR model), we would have only identified 530,186 additional debt-related articles which when combined with journalist labels would have increased our dataset size by only 17%. Had we only taken articles identified as debt-related by the DMR model, our final dataset would only contain 1.5 million (1,528,266) news articles, a dataset that has half as many observations as the baseline dataset. In addition to increasing our dataset in size by 54%, we also construct a dataset that corrects for concept drift and reflects current notions of what counts as debt-related news. This will impact downstream research, the predictive power of future models, and any other findings that use this debt-related news dataset. It is plausible that in many use cases, the viability of a model or research question hinges on how effectively a forecaster is able to address inconsistent labeling.

6. Conclusion

We consider the problem of inconsistent labeling in text datasets and propose a solution that synthesizes attribution and prediction methods. An integral part of our labeling solution is the selection of an accurate text classification model, and we consider the tradeoffs between attribution-based models and pure prediction models in the context of labeling 27 million

Thomson Reuters news articles. We hope to impress the following findings upon the reader. Issues of missing labels, inconsistent labels, or concept drift are likely to plague any dataset with human labeling, especially if the dataset covers a long period of history. These labeling issues must be addressed, and even if the forecaster does not suspect inconsistent labeling or concept drift they should test for the existence of these problems. Using our proposed solution, a forecaster may both test for these problems and solve them by backfilling labels with a classification model. However, when implementing this backfill strategy and in general, the tradeoffs between attribution and prediction are material and should be thoroughly considered instead of decided as a matter of taste or convenience. While we focus on text analysis, problems of inconsistent labeling are likely to occur in many domains ranging from fraud detection to image classification. We demonstrate how by combining insights from attribution methods with the power of pure prediction models, these issues may be overcome and the value of inconsistently labeled datasets unlocked. We hope that our work motivates practitioners of both modeling methods to pay closer attention to methods that are not their own and develop new ways to synthesize these methods to achieve better forecasting.

References

Athey, S., et al. (2021), "Integrating explanation and prediction in computational social science", *Science*, 595, 181-188.

Baker, S., Bloom, N., and Davis, S. (2016). "Measuring Economic Policy Uncertainty", *The Quarterly Journal of Economics* 131, 1593-1636.

Barbaglia, L., Consoli, S., and Manzan, S. (2022), "Forecasting with Economic News", *Journal of Business & Economic Statistics.*

Breiman, L. (2001). "Random Forests", *Machine Learning,* 45**,** 5-32.

Brown, T., et al. (2020), "Language Models are Few-Shot Learners".

Chen, T., Guestrin, C. (2016). "XGBoost: A Scalable Tree Boosting System".

Devlin, J., Chang M., Lee K., Toutanova K. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding".

Efron, B. (2020). "Prediction, Estimation, and Attribution", *Journal of the American Statistical Association*, 115, 636-655.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). "Text as Data", *Journal of Economic Literature*, 57, 535-74.

Hansen, S., McMahon, M., and Pratt, A. (2018). "Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach", *The Quarterly Journal of Economics*, 133, 801-870.

Helmbold, D., Long, P. (1994). “Tracking drifting concepts by minimizing disagreement”, *Machine Learning*, 14, 27-46.

Laver, M., Benoit, K., & Garry, J. (2003). “Extracting Policy Positions from Political Texts Using Words as Data”, *American Political Science Review*, 97, 311-331.

Londono, J., Claessens S., and Correa, R. (2021). “Financial Stability Governance and Central Bank Communications,” *International Finance Discussion Papers 1328. Washington: Board of Governors of the Federal Reserve System*.

Loughran, T., McDonald, B. (2011). “When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks”, *The Journal of Finance*, 66, 35-65.

Rumelhart, D., Hinton, G., and Williams, R. (1986). “Learning representations by back propagating errors”, *Nature*, 323, 533–536.

Ryland Williams, et al. (2015). “Zipf’s law holds for phrases, not words”, *Scientific Reports*, 5, 1-7.

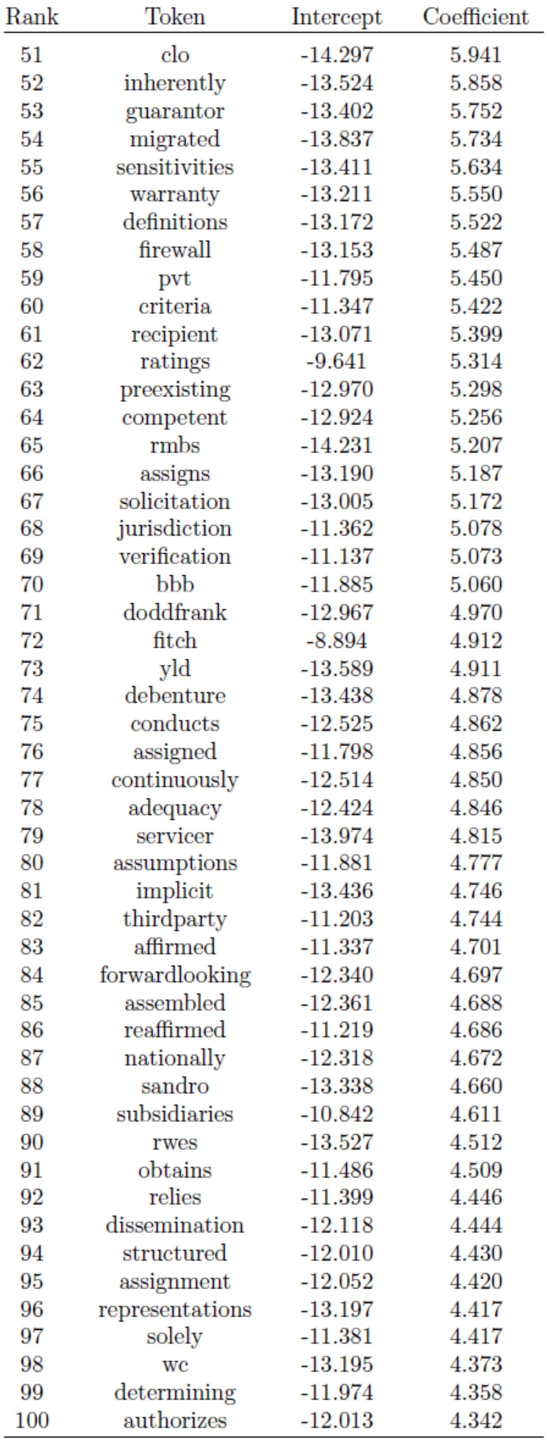
Taddy, M. (2013). “Multinomial Inverse Regression for Text Analysis”, *Journal of the American Statistical Association*, 108, 755-770.

Taddy, M. (2015). “Distributed Multinomial Regression”, *The Annals of Applied Statistics*, 9, 1394-1414.

Werbos, P. (1990). “Backpropagation through time: what it does and how to do it,” *Proceedings of the IEEE*, 78, 1550-1560.

Zipf, G. K. (1935). “The Psycho-Biology of Language”, *Houghton-Mifflin*.

Appendix 1: DMR Coefficients 51 - 100



Rank	Token	Intercept	Coefficient
51	clo	-14.297	5.941
52	inherently	-13.524	5.858
53	guarantor	-13.402	5.752
54	migrated	-13.837	5.734
55	sensitivities	-13.411	5.634
56	warranty	-13.211	5.550
57	definitions	-13.172	5.522
58	firewall	-13.153	5.487
59	pvt	-11.795	5.450
60	criteria	-11.347	5.422
61	recipient	-13.071	5.399
62	ratings	-9.641	5.314
63	preexisting	-12.970	5.298
64	competent	-12.924	5.256
65	rmbs	-14.231	5.207
66	assigns	-13.190	5.187
67	solicitation	-13.005	5.172
68	jurisdiction	-11.362	5.078
69	verification	-11.137	5.073
70	bbb	-11.885	5.060
71	doddfrank	-12.967	4.970
72	fitch	-8.894	4.912
73	yld	-13.589	4.911
74	debenture	-13.438	4.878
75	conducts	-12.525	4.862
76	assigned	-11.798	4.856
77	continuously	-12.514	4.850
78	adequacy	-12.424	4.846
79	servicer	-13.974	4.815
80	assumptions	-11.881	4.777
81	implicit	-13.436	4.746
82	thirdparty	-11.203	4.744
83	affirmed	-11.337	4.701
84	forwardlooking	-12.340	4.697
85	assembled	-12.361	4.688
86	reaffirmed	-11.219	4.686
87	nationally	-12.318	4.672
88	sandro	-13.338	4.660
89	subsidiaries	-10.842	4.611
90	rwes	-13.527	4.512
91	obtains	-11.486	4.509
92	relies	-11.399	4.446
93	dissemination	-12.118	4.444
94	structured	-12.010	4.430
95	assignment	-12.052	4.420
96	representations	-13.197	4.417
97	solely	-11.381	4.417
98	wc	-13.195	4.373
99	determining	-11.974	4.358
100	authorizes	-12.013	4.342

Note: Top 51-100 token coefficients estimated by the DMR model. The presence of these tokens in a news article is likely to result in the news article being classified as debt-related. These tokens further indicate that the DMR model is using sensible features of the text data to classify articles as debt-related or non-debt-related.

Appendix 2: Hyperparameters

DMR Model (R package: distrom, function: dmr)

mu = NULL

bins = NULL

gamma = 1

nlambda = 50

Feedforward Neural Network (Python package: PyTorch)

nn.Linear(5000, 100)

torch.relu()

nn.Linear(100, 10)

torch.relu()

nn.Linear(10, 1)

torch.sigmoid()

batch_size = 10

epochs = 5

learning_rate = 0.01

loss_function = torch.nn.MSELoss()

optimizer = torch.optim.SGD()

BERT Model (Python package: transformers, function: DistilBertForSequenceClassification)

100 token input.

6 transformer blocks of feature size of 768

768 neuron pre-classification laye

2 neuron output layer

tokenizer = transformers.BertTokenizerFast()

batch_size = 10

epochs = 1

learning_rate = 1e-5

optimizer = transformers.AdamW()

Random Forest (R package: ranger, function: ranger)

num.trees = 100

mtry = 70

min.node.size = 1

max.depth = NULL

sample.fraction = 1

splitrule = gini

XGBoost

objective = binary:logistic

eval_metric = auc

eta = 0.3

max_depth = 6

min_child_weight = 1

subsample = 1

colsample_bytree = 1

lambda = 1

alpha = 0

nrounds = 10000