How Informative Is the Text of Securities Complaints?

Adam B. Badawi*

November 2021

Abstract

Much of the research in law and finance reduces long, complex texts down to a small number of variables. Examples include the coding of corporate charters as an entrenchment index or characterizing dense securities complaints by using variables that capture the amount at issue, the statutes alleged to have been violated, and the presence of an SEC investigation. Legal scholars have often voiced concerns that this type of dimensionality reduction loses much of the nuance and detail that is embedded in legal text. This paper assesses this critique by asking whether methods that can analyze text are able to capture meaningful—and perhaps even more—information than traditional low-dimension studies that rely on non-textual inputs. It does so by applying text analysis and machine learning to a corpus of more than five thousand complaints filed in private securities class actions that collectively contain over 90 million words. This analysis shows that there is significant information embedded in the text of these complaints, albeit with substantial limitations on how much information that text analysis can extract. The analysis proceeds in three parts. The first asks whether the text provides indications about the eventual outcomes in the cases. The best performing models predict whether cases will settle or get dismissed with an accuracy rate of about 70 percent. That is substantially better than baseline rates, but still leaves significant room for improvement. The second part of the analysis compares text-based models to non-text models and assesses their relative performance in predicting outcomes. While the best performing text-based models are more accurate than the best performing non-text models, a hybrid model that uses both text and non-text components performs better than either of these two components alone. These results suggest that there may be some information omitted

*Professor of Law, UC Berkeley School of Law. I thank Joe Grundfest for providing access to data from the Stanford Securities Class Action Clearinghouse. I also thank Elliott Ash, Ken Ayotte, Bobby Bartlett, Steve Choi, Jens Frankenreiter, Jonah Gelbach, Craig Lewis, Josh Mitts, Adair Morse, Anthony Niblett, Julian Nyarko, Michael Ohlrogge, Gabe Rauterberg, Adriana Robertson, Jonathan Slapin, Eric Talley, Kate Waldock and audiences at the Vanderbilt Law and Business Conference, the Zürich Text as Data Conference, the Berkeley Text Analysis in Law Conference, the Corporate Law Academic Workshop Series, the Max Planck Institute for Research on Collective Goods, the University of Amsterdam, the University of Toronto, and Erasmus University for helpful comments.

from the non-text models and that augmenting them with textual information may improve them. Finally, the analysis uses abnormal returns as an additional means of validation. Previous research shows that there are substantial differences in the abnormal returns of cases that will get dismissed and those that will settle in the days following the filing of a securities lawsuit. This section replicates this result and then shows that the predictions made by the machine learning models are associated with substantial abnormal returns. While market participants take about three or four days to settle on the likely outcome of a case on stock price, the machine learning models can make these predictions more or less instantaneously. In addition, to validating the predictions against human judgment, these results also suggest that there is some stock price drift in the reactions to the complexities of securities lawsuits.

1 Introduction

It has long been a practice in law and finance to reduce long, complex pieces of text—such as statutes, court opinions, and corporate charters—down to a single or, maybe, a handful of variables. Examples of this approach are widespread and include research into takeover defenses (Gompers, Ishii, and Metrick 2003; Bebchuk, Cohen, and Ferrell 2008), securities litigation (Choi and Pritchard 2016), and even entire legal regimes (La Porta et al. 1998). While reducing dimensionality is a necessary part of understanding complex phenomena, there have been persistent critiques by legal scholars that coding variables in too narrow a way misses important information that is embedded in the relevant legal texts. For example, Catan and Kahan (2016) argue that coding of corporate charters misses important legal nuances in takeover law and Spamann (2009) shows that close attention to the underlying legal text casts doubt on some of the primary conclusions of the legal origins literature. This paper asks whether using machine learning and text analysis—which may be able to pick up some of the details that the critics identify–produces meaningfully different results from low-dimension, non-text inputs. It does so by analyzing whether using these techniques on the the text of several thousand private securities complaints provides useful information about the outcomes in those cases and compares that text-based analysis to predictions made by non-text variables and to the judgments of stock traders.

This text-focused analysis proceeds in three stages. The first stage assesses whether the text in the complaint provides information about the eventual success or failure of each case. The machine learning models that rely only on the textual content of the complaints are able to predict whether the first-filed complaints and the consolidated complaints (i.e. those filed after the selection of lead counsel) get dismissed or settle at rates approaching 70 percent. These accuracy rates are well above the baseline rates and provide evidence that there is information in the text that indicates what is likely to happen in the case. This text may capture traditional measures of case quality—such as the statutory basis for the claim or whether the SEC investigated the matter—but it may also be capturing more subtle information such as the presence of verifiable "hard facts" that indicate a stronger case. The machine learning techniques allow for some insight into these potential explanations and provide modest support for the "hard facts" theory. While the above-baseline performance of these text-based models is noteworthy, an accuracy rate around 70 percent leaves substantial room for improvement.

The second stage compares the performance of text-based models to their non-text counterparts. The best performing text-only models from the first stage outperform all of the non-text models evaluated in the second stage, albeit by modest amounts. Those non-text models use standard variables from securities research such as the statutory basis of the claim, the court of filing, the industry of the defendant firm and the loss from the alleged fraud for first-filed cases. But when the text models are modified in a way that allows them to be combined with the non-text variables, the differences between these limited text-only models and the non-text-only are marginal. Combining the two sources of data produces higher accuracy rates than either the modified text-only models or non-text-only models on their own when classifying the first-filed complaints. Yet these combined models still under perform the best versions of the text-only models. Together, these results show that there is some extractable information in the text that does not get fully captured by the non-text variables used in these analyses.

The final stage examines how the speed and accuracy of machine learning models compare to the judgments of stock traders. Previous research shows that market participants are able to identify the likely outcome in securities class actions, albeit with some delay (Griffin, Grundfest, and Perino 2004). This phenomenon, which I replicate with the newer data that I collect, allows abnormal returns to serve as another way to validate the predictions of the machine learning models. To do so, I construct a portfolio that goes long on the quintile of cases that the model predicts are most likely to settle and goes short on the quintile of cases that the model predicts are most likely to get dismissed. This portfolio is associated with substantial abnormal returns and those returns fall when I include a broader range of predictions (e.g. the tercile of cases most likely to settle and the tercile of cases least likely to settle). These results provide evidence that predictions based on the text alone can identify not only the strong and weak cases, but the relative strength and weakness of those cases. The abnormal returns are, however, a temporary phenomenon. After about four days, the returns produced by the predictions no longer earn outsized returns.

These findings have methodological and substantive implications. On methodology, the evidence that the text is informative and that text-based models modestly outperform non-text-based models raises potential concerns about omitted variable bias. Most studies of business litigation use non-text variables to control for case quality. In research on securities litigation, these controls often include the size and industry of the defendant company, the alleged damages, whether the SEC investigated the allegations, and the statute underlying the claim. Analyzing the text of the complaints may provide another way to capture non-text variables (e.g., the intensity of an SEC investigation) and it may register variation that is not easily reduced to a non-text variable (e.g., the presence of hard facts). And, indeed, this paper shows that the best performing text-only models outperform models that only use non-text variables, albeit by small amounts. These results suggest that, where possible, researchers should use both text and non-text variables when accounting for variation between cases and that failing to include text-based variables risks introducing omitted variable bias into securities litigation research. The results also suggest that solving this problem is not trivial. Combining text and non-text models often requires further reduction of the dimensionality of text information and that naturally degrades the performance of the text models. Going forward, an important challenge for law and finance researchers will be to develop methods to assess the amount of potentially omitted information that legal text contains and to find ways to combine that information with non-text variables in a way that can better account for that omitted information.

Substantively, the results contribute to the literature on stock price drift when reacting to complex information. While the precise magnitude of the abnormal returns varies based on the timing of the trades and the strength of the predictions in the underlying portfolio, there is strong evidence that the machine learning predictions are associated with abnormal returns. This effect is, however, temporary. Market prices appear to absorb the information in the complaint within several trading days. As with Cohen, Malloy, and Nguyen (2020), these results suggest that market participants take time to process complex information that, at least within the time frame of this study, text analysis and machine learning techniques can pick up more quickly. More generally, these results join other studies that use text analysis to uncover pricing anomalies in financial markets (Edmans, Garcia, and Norli 2007; Garcia 2013).

The results also help to corroborate some conventional wisdom among securities law practitioners and commentators. The machine learning models better predict outcomes for first-filed complaints than they do for the consolidated complaints that get filed after the selection of lead counsel. This result is consistent with the distribution in quality of these complaints being greater for first-filed complaints than for consolidated complaints. If this interpretation is correct, it would help to confirm the perception that the Private Securities Litigation Reform Act of 1995 ("PSLRA") has not totally eliminated the race to the courthouse and that competition for lead plaintiff status has helped to produce higher quality work product (Weiss 2008).

This paper proceeds as follows. Section 2 reviews the relevant literature on securities litigation, lawyering, and the use of text analysis in law and finance. Section 3 reviews data collection and the construction of the text corpus. Section 4 provides an overview of the machine learning models and presents the results of the attempt to classify securities lawsuit outcomes based on the textual and non-textual features of the case. This section also examines whether machine learning predictions are associated with abnormal returns in the stock market. Section 5 concludes.

2 Literature Review

This project draws on the literatures on securities litigation, the relationship between lawyer quality and case outcomes, and the use of text analysis and machine learning in law and finance. This section reviews each of these literatures in turn.

2.1 Private Securities Enforcement and the PSLRA

Private plaintiffs may bring a securities fraud class action under several federal statutes. The most commonly used statute for this purpose is Section 10b of the Securities Act of 1934 (and the associated SEC Rule 10b-5). In response to what was perceived to be excessive private securities litigation, Congress enacted the Private Securities Litigation Reform Act of 1995, which continues to control the structure and process of most private securities litigation.[1] The PSLRA imposes, among other requirements, a heightened pleading standard relative to other litigation and it requires a lead plaintiff selection process that gives priority to class members who hold the largest financial interest in the case. The heightened pleading standard in the statute requires plaintiffs to plead "with

[1] It is worth noting that a meaningful amount of private securities litigation proceeds under Section 11 of the Securities Act of 1933. These cases allege a material misstatement in connection with a securities offering (typically an initial public offering) and do not require a showing of intent. For this reason, the most important of the heightened pleading requirements of the PSLRA do not apply to Section 11 actions (Choi 2006).

particularity facts giving rise to a strong inference that the defendant acted with [scienter]." The lead plaintiff procedures sought to undermine incentives to race to the courthouse by awarding lead plaintiff status to those stockholders with the largest financial interest in the case (typically institutional investors).

Assuming that the largest investors are willing to get involved in securities litigation and assuming that these investors can identify the better lawyers, these changes should result in higher quality complaints in the post-PSLRA world. This higher quality should result both from the heightened pleading standard and the work of better lawyers. This effect should be especially evident for the consolidated complaints that get filed after the selection of a lead plaintiff. But these assumptions depend on the PSLRA achieving its policy goals and the empirical evidence is quite mixed on this point. For example, Cox and Thomas (2006) find some increase in the number of public institutional investors after the PSLRA, but they find few differences before and after the legislation with respect to settlements and defendant characteristics. The authors also find that the ratio of settlement amount to provable losses actually declined after the PSLRA. Choi, Nelson, and Pritchard (2009) examine the pre and post-PSLRA periods and find that there has not been a detectable decrease in the number of nuisance suits that get filed, but there has been an increase in the number of meritorious suits that get screened out. Choi, Fisch, and Pritchard (2005) conduct a pre and post analysis of lead plaintiffs and find that the presence of private institutional plaintiffs is not associated with higher settlements after the PSLRA, but there is some evidence of higher settlements when public institutional investors act as lead plaintiffs. They also find that institutional investors are not associated with lower fee awards to plaintiffs' attorneys. More anecdotal accounts suggest that the PSLRA did not do much to prevent the race to the courthouse that it sought to prevent (Weiss 2008). The data I use in this study contain only post-PLSRA cases and thus I cannot make before and after comparisons. I can, however, distinguish between first-filed and consolidated cases and I can assess the post-PSLRA differences in these two types of complaints.

Scholars have examined the relationship between securities class action characteristics and their consequences for firms. A substantial amount of this research looks at the stock price reaction to events related to the litigation. Pritchard and Ferris (2001) find substantial negative abnormal returns associated with the disclosure of the alleged fraud and a smaller negative return in the three-day window surrounding the filing of the complaint. Griffin, Grundfest, and Perino (2004) also find negative stock price reactions associated with the disclosure of corrective information and class action filing. They show that stock prices react differently to filing depending on the ultimate outcome of the case and that market has a more negative reaction to cases that allege accounting deficiencies. Choi and Pritchard (2016) compare cases that involve an SEC investigation but no class action to those that involve a class action but no SEC investigation. They show that the class action-only cases are associated with larger stock market drops, increased institutional ownership turnover, and more and larger settlements.

2.2 Lawyer Quality and Lawsuit Merit

This study examines whether the words that the lawyers choose to put in the complaint can be used to predict whether a securities class action. This aspect of the study ties into recent research on measuring lawyer and lawsuit quality and the associations between that quality and outcomes. Badawi and Webber (2015) show that merger challenges filed by law firms that rank highly along a number of dimensions are associated with higher abnormal stock market returns. Krishnan, Solomon, and Thomas (2015) develop a measure of plaintiff law firm quality that takes into account their representation of informed clients and their ability to obtain large settlements. They show that these law firms are more successful in their cases and that they file more documents than lesser quality law firms. In another study, the same authors identify high quality defense counsel based on league tables (Krishnan, Solomon, and Thomas 2017). They find that these firms are associated with more favorable outcomes for their clients. None of these studies, however, focus on the content of the documents produced by lawyers to assess the quality of the lawyers or a lawsuit.

Determining the merit of a lawsuit is a difficult empirical task. Nevertheless, scholars in different areas of law and finance have developed some ways to control for the quality of lawsuits. In research on securities litigation, it is relatively common to control for whether the alleged conduct underlying a case was investigated by the SEC, the amount of alleged damages, and defendant characteristics such as industry and market capitalization (see, e.g., Griffin et al. 2008). In the context of merger and derivative litigation, researchers have used variables such as transaction size, type of merger consideration, structure of a transaction, and deal premium to control for case quality (Cain and Solomon 2014). Some studies also use the number of cases filed as a way to capture case quality (Cain and Solomon, (2015); Badawi and Chen (2017)). But, to date, there do not appear to be any studies that use the text of case documents to generate measures of case merit.

2.3 Text Analysis in Law and Finance

Both the legal and finance literatures have made substantial use of text analysis and, to some degree, machine learning. The legal literature has largely focused on prediction of outcomes, but has done so mostly in the context of using the text of legal opinions to predict the outcomes of those very opinions. In finance, the literature has focused on using text to generate data sets that can be used in service of conventional analysis and on using text analysis to unearth pricing anomalies in the stock market. I discuss these two areas in turn.

The emerging field of computational law has mostly, although not exclusively, focused on classification and prediction rather than explanation. Earlier work sought to identify the topics contained in legal text (Gonçalves and Quaresma 2005) and classifying argumentative versus non-argumentative text (Palau and Moens 2009). More recent scholarship seeks to classify ex post litigation outcomes. Several of these projects use the text of court opinions to predict the outcomes of those opinions. For example, Aletras et al. (2016) use the text of opinions from the European Court of Human Rights to predict the outcome of those cases. They predict the outcome with 79% accuracy and find that the discussion of facts of in a case are the most important factor for prediction. Sulea et al. (2017) use the text of case descriptions from the French Supreme Court to predict the outcomes of those decisions with about 97% accuracy.

There are fewer projects that, like this one, use ex ante information to predict outcomes and it does not appear that any other other papers use ex ante text produced by an adversarial party to predict litigation outcomes. Examples of these ex ante approaches include Katz, Bommarito II, and Blackman (2017), which uses non-text features to predict US Supreme Court outcomes with 70 percent accuracy and Wongchaisuwat, Klabjan, and McGinnis (2017), which predicts the likelihood of a patent being litigated based on the text of the patent and several non-text features. Their highest performing model obtains an f-1 score of 19 percent against a baseline litigation rate of one to two percent.

In finance, there have been two dominant uses of text analysis and machine learning. The first is to use text—largely from EDGAR—to develop data sets that can be used in conventional analysis. Examples in this genre include Hanley and Hoberg (2010), which provides evidence of a relationship between the amount of individual tailoring in a registration statement and the amount of IPO underpricing and Hoberg and Phillips (2016), which uses text analysis of product descriptions to reclassify industries based on competitors with similar products. Some work combines text analysis with machine learning, such as Buehlmaier and Whited (2018), who use a Naive Bayes algorithm on the text of management's discussion and analysis in disclosures to identify financially constrained firms. Others have used machine learning techniques to identify trends in corporate governance (Rauterberg and Talley 2017) and contracting (Nyarko 2021).

The other main use of text analysis has been to identify instances where market participants have been slow to identify information that affects stock prices. Cohen et al. (2020) show that the stock market is slow to notice changes in the text of the quarterly filings and annual reports of public firms. A portfolio that shorts the firms that alter this language the most and goes long on those that change it the least earns up to 188 basis points in monthly alpha. Garcia (2013) documents that sentiment, as expressed in financial news in the New York Times, is predictive of stock returns during recessions. The analysis in this paper follows this second approach by demonstrating that predictions based on the text of securities complaints is associated with substantial abnormal returns. But the predictions generated based on complaint text could also be used in support of more conventional analysis. For example, those predictions could serve as a control for the textual merit of a case in studies of securities litigation.

3 Data and Corpus Construction

The data and text for this project come from the Stanford Securities Class Action Clearinghouse (SCAC). This database provides a comprehensive overview of securities filings in the United States going back to 1996. At the time of data collection, SCAC contained information on over 4,000 cases. The database provides both structured data about cases as well as important documents. The structured data includes case status (settled, dismissed, or ongoing), the federal district of the case, and names of the plaintiffs' firms litigating the case. The important documents include the first-filed complaint, the consolidated complaint, the docket sheet, major rulings, and the details of any settlement, if there is one. Using Python, I scrape all the structured data as well as the first-filed complaint and the consolidated complaint from every case in the database as of late 2019. After excluding ongoing cases and eliminating cases where the documents do not appear to be complaints, the complete data set of first-filed cases includes 3386 observations and the complete data set of consolidated cases includes 2382 observations. I describe the processing of the text in the Technical Appendix.

The machine learning analysis applies an algorithm to the processed text to predict whether a securities case is dismissed or settled. To assess performance, I use a simple measure of accuracy (the number of correct predictions divided by the total number of predictions), which is appropriate in applications such as this where the outcomes of interest—settlements and dismissals—are not rare. In all the analyses, I perform 100 iterations where each iteration randomly selects 90 percent of the observations to build a model that predicts the remaining 10 percent of the observations. I report the average accuracy of those predictions and scores at the 2.5 percentile and 97.5 percentile.

Some of the analysis uses non-textual data. I obtain some of this information from the SCAC database, which includes filing dates, the court where plaintiffs file the complaint, SIC industry classification, and whether the complaint presents claims under Section 10(b) of the Securities Exchange Act of 1934 and/or Section 11 of the Securities Exchange Act of 1933. I also obtain the disclosure dollar loss associated with each case where the disclosure dollar loss is difference in market capitalization for the defendant firm on the trading day prior to the last day in the class period and the trading day on the day after that date as reported by the Center for Research on Security Prices (CRSP). This information is not available for all of the cases. After merging together these data sets, there are 2702 first-filed cases and 1989 consolidated cases.

Table 1 shows some basic statistics for the first-filed cases. The year of filing ranges from 1996 to 2019 with a median of 2008. Over three-quarters of the cases involve a claim under Section 10b and a little under 14 percent of the cases involve a Section 11 claim. About 44 percent of the cases settle and almost three-quarters of the cases are eventually consolidated. A little over a 23 percent of the cases get filed in a New York federal court, with the vast majority of these cases getting filed in the Southern District of New York. About 21 percent of cases get filed in a California federal court, with the majority of these cases getting filed in the Northern District of California. Nearly 28 percent of the cases filed involve a defendant in the technology industry.[2] Finally, and as one would expect, consolidated complaints are longer, on average, than first-filed complaints. First-filed complaints average 9150 words and the average consolidated complaint approaches three times that length.

[2]I define technology firms as those with an SIC in any of the following groups: Biotechnology & Drugs, Computer Hardware, Computer Networks, Computer Peripherals, Computer Services, Semiconductors, or Software & Programming.

4 Analysis and Discussion

This section provides the details and results of the classification exercise. The first subsection provides an overview of the machine learning techniques that I use and then discusses the results of an analysis that uses only the text of the first-filed and consolidated complaints to predict whether a securities class action will settle or will be dismissed.[3] The next subsection incorporates non-text features into the analysis, such as the year of filing, the identity of the federal district court, and the type of claim to determine whether including this information improves prediction. The final subsection examines the relationship between the machine learning predictions and the stock prices of defendant firms around the filing of the complaints.

4.1 Predicting Case Outcomes on the Basis of Complaint Text

The process for outcome prediction follows a typical machine learning classification process. The cleaned data, which in most cases is a matrix of tf-idf weighted terms, is run through different sets of predictive algorithms. This task is an instance of supervised learning because the variable of interest—case outcome—is incorporated into the building of the model.[4] Each of these algorithms fits a model to the in-sample data, which can then be used to generate predictions for out-of-sample data (i.e. the ten percent of the data that has been randomly selected to be held out). Explanations of the individual algorithms that I use are in the methodological appendix.

Table 2 presents the results for classifying based on text alone. Panel A provides the outcomes for an analysis of the first-filed complaints in each case, and Panel B provides the results for the consolidated complaint in cases that reach that stage. To assess the performance of the classification exercise it is helpful to have a baseline. One is simply chance. Are the predictions better than flipping a coin to predict whether a case gets dismissed or is settled? Perhaps a better benchmark would be to observe whether the classification can improve on guessing the modal outcome. For the first-filed complaints, dismissal is the modal outcome, which occurs 56.3% of the time. For the sample of

[3] The vast majority of securities class actions are either dismissed or settled. See https://www.cornerstone.com/Publications/Reports/Securities-Class-Action-Filings-2018-Year-in-Review. Typically, the dismissals are either voluntarily agreed to by the parties or are the consequence of a court granting a motion to dismiss.

[4] In unsupervised learning, the algorithm is not provided information about the outcome of interest. Instead, the researcher will typically specify a number of groups to partition the data into and the algorithm will attempt to find the best way to group the data in this way.

consolidated complaints, the modal case settles and does so 56.7% of the time. Across all the complaints the cases get dismissed 51.0% of the time.

Each of the panels in Table 2 provides the mean accuracy rate and the 2.5 and 97.5 percentile accuracy rates. In Panel A, the random forest classifier performs best followed by the extra trees model and the Adaboost classifier. Both of these algorithms correctly classify whether a securities complaint is dismissed or settled about 70 percent of the time. This prediction is a substantial improvement over the baseline of dismissal in 56.3% of the cases. The XGBoost classifier is just behind the top two algorithms and the Naive Bayes algorithm is a few percentage points behind.

The results are somewhat different for the consolidated complaints in Panel B. The top classifiers for these complaints are the XGBoost and extra trees classifiers, which correctly classify outcomes 65.7% and 65.4% of the time respectively. The Naive Bayes algorithm performs the worst on this task.[5] These results are against a baseline of 56.7% of the cases settling and are obtained in the context of cases where a court has selected counsel in a competitive process and where those counsel typically devote significant resources to drafting the consolidated complaint.

These results have several implications. The first is simply that the text is informative for both first-filed and consolidated complaints. In both sets of analysis, all of the algorithms perform better than the baseline projections. To the degree that dismissals and settlements are a gauge of lawsuit quality, the words that the lawyers use in the complaints say something about the quality of the lawsuit. This analysis cannot, however, differentiate between what the words chosen say about the quality of the lawyers bringing the case and the underlying merits of the case. It could be that more skilled lawyers describe cases in a way that reflects that skill and the text analysis is picking up these choices. It may also be the case that the words reflect the type of case and that some types of cases are more likely to succeed than others. The terms that are most helpful in discriminating between outcomes, which are provided in Table A2 of the Appendix, provide some indication that the timing and topic of the case are useful in making predictions.

A natural question is why is the text informative? While a common critique of machine learning models is that they are black boxes that provide little opportunity for interpretation or explanation, it is possible to provide some insight to how the models are behaving by looking at simplified versions of them. Figure 1 shows one of the decision

[5] The weaker performance of the classifiers on the consolidated complaints could be due to the reduction in the number of observations. To rule out this possibility, I rerun the classification exercise on 2382 randomly selected first-filed complaints. The results of this exercise are nearly identical to those in Panel A of Table 2, which includes the full sample of first-filed complaints.

trees in a random forest model that I limit to three leaves.[6] The most important term for discriminating between settled and dismissed cases is "statements false," with a tf-idf weight cutoff of .01. Complaints where this weight is higher (i.e., the term is used more frequently) tend to be associated with settlement. Where a complaint uses this term frequently, the next most important term is "quarters," where cases again benefit from more frequent use of this term. The straightforward implication is that cases that involve a larger number of false statements over longer periods of time tend to be stronger than those that do not. For those cases to the left of the tree—those that involve fewer uses of the term "statement false"—the next most important term for classification is "10b," which suggests that Section 10b cases are more likely to be successful than others. A likely reason for this is that Section 14 cases, which mostly involve challenges to mergers under the securities laws that rarely produce settlements (Cain et al. 2019). This interpretation finds support in Table A2, which shows that the terms "14a" and "section 14" are especially good at distinguishing cases. It is also worth noting that the common trope that lower quality securities lawyers rely on news reports for their research finds support in that lesser use of the term "media" is associated with settlement.

This exercise suggests that one advantage of using text-based models to analyze securities complaints is that they identify subtle differences in the cases that might be difficult or resource intensive to code. The analysis presented in Figure 1 suggests that some of these subtleties relate to allegations of "hard" facts—such as provably false statements—rather than "soft" facts that suggest only generalized allegations of fraud. The text-based models may be able to identify the relative importance of terms associated with these subtle differences and the frequency with which those terms appear. Both of these tasks might prove to be a challenge for human coders who wish to create a set of useful non-text variables. Those coders might be able to specify the number of alleged false statements in a complaint or the duration of the alleged wrongdoing—although inter-rater reliability might be a challenge—but one suspects it would be especially difficult for human coders to read a corpus of complaints and identify the terms or facts that best distinguish the quality of those complaints. Machine learning and text analysis have an advantage at making these sorts of distinctions.

A second implication of these results is that it appears to be easier to classify based on first-filed complaints rather than consolidated complaints, especially when taking into account the baseline results for these cases. This result suggests that there is broader

[6] In the complete model, there is no limit on the number of leaves. The model will use as many leaves as is necessary to perfectly discriminate between all the complaints.

variation in the first-filed complaints than there is in the consolidated complaints. Broader variation would make the classification task easier for the first-filed complaints and narrower variation would make the classification more difficult for consolidated complaints. This pattern is consistent with the conventional wisdom about the post-PSLRA world. While the PSLRA may have put some brakes on the race-to-the-courthouse, some practitioners believe that it still occurs.[7] Narrower variation in the quality of consolidated complaints would be consistent with the lead plaintiff provisions of the PSLRA having some of its the intended effect. If plaintiffs with a larger financial interest sign up to be lead plaintiffs and if they choose high quality counsel, one would expect the work product produced by the lawyers to be relatively high. This pattern, if widely adopted, could produce relatively little variation in the quality consolidated complaints. But it is important to emphasize that the evidence is only consistent with these accounts. These models cannot make absolute quality comparisons between first-filed and consolidated complaints and it cannot make any comparisons between pre and post PSLRA cases.

4.2 Classification Robustness Checks

There are several methodological and substantive concerns about the robustness of the results presented in the last section. The primary methodological concern is the use of hyperparmeter tuning. The results in Table 2 reflect choices about the number of features to include, the number of n-grams, and the number of estimators for some of the models. Those choices reflect observed performance. There is some potential that using observed performance to tune the model will produce overly optimistic classifications because this approach does not preserve a complete separation between training data and test data. While simulations suggest that this concern is overblown in many applied settings, nested cross validation is a solution to this problem (Wainer and Cawley 2018), albeit a computationally expensive one (Cawley and Talbot 2010). To assess the potential for overly optimistic estimates, I use this cross-fold validation on some of the models used in Table 2. I describe nested cross-validation and the parameters used in the robustness checks in the Technical Appendix.

Table A1 presents the mean, 2.5 percentile, and 97.5 percentile for the 100 nested cross validation estimates of random forest and extra trees models for both the first-filed complaints and the consolidated complaints. While all of the estimates are slightly lower than those from Table 2, they are within a percentage point of the Table 2 estimates. In

[7] Weiss (2008) argues that some plaintiffs' firms may still file cases quickly because having filed a case makes it easier to recruit additional class members.

most cases the differences are within a few tenths of a percentage point. These results suggest that the data leakage that comes with hyperparameter tuning produces estimates that are only slightly higher than those produced by nested cross fold validation. The more important point from this exercise is that even with strong insulation against any data leakage, the models produce accuracy rates that are substantially above baseline settlement and dismissal rates.

Substantively, there are three primary concerns that may lead to estimates that are overly optimistic. The first is that the most recent cases in the dataset are far more likely to be dismissed than to be settled. A case that produces a securities settlement will typically take several years to be finalized while dismissals can happen as quickly as a few weeks. There is a danger that the classification models will use indications of recency to predict outcomes rather than the substantive content of the complaints. There are indications of this issue in Table A2, which lists the most important words for classification. The words "cv" and "case" are contained in the header of electronically-filed cases. Insofar as more recent cases are likely to get filed electronically, the models may be focusing on these terms because they are associated with the recent cases that produced dismissals.

I run the classification models on a dataset that omits all cases from 2017 to the end of the sample in 2019. Doing so naturally changes the baseline outcomes for the cases. In this subsample 51.0% of the first-filed cases settle, which is in contrast to 56.3% of the cases getting dismissed in the full sample. In 100 iterations of the Extra Trees classifier on the subsample, the average accuracy rate is 65.1% and for 100 iterations of the Random Forest classifier the mean accuracy rate is 64.0%. Both of these numbers are substantial improvements on the 51.0% baseline and suggest that, even without the recency clues, the text of the first-filed complaints provides insight into the ultimate outcomes of the cases.

A second substantive concern is the sharp increase in challenges to mergers being brought under the federal securities laws. As discussed above, these cases often result in dismissals. If there are terms that indicate these types of suits, the ability to identify these suits may be a substantial reason for a classifier's accuracy. Merger lawsuits under the securities laws are typically brought under Section 14(a) of the Exchange Act of 1934. The appearance of the terms "14a" and "Section 14" thus indicate that these terms are helpful for classification. To determine how well the models are able to classify cases without the inclusion of merger cases I use data supplied by SCAC to screen out all cases where the only claims are based on Section 14(a) or Section 14(e).

Without the merger cases the mean accuracy for 100 iterations of the Extra Trees classifier is 64.0% and the mean accuracy of the Random Forest classifier is 63.9% for the first-filed complaints. These are sizable drops from the rates in the 69-70% range for models that include the merger cases. But, as with excluding the recent cases, excluding the merger cases changes the baseline. The baseline for the non-merger cases is 52.3% of those cases getting settled. Even when omitting the cases that are easier to classify, the models are still able to classify at accuracy rates that are substantially above the baseline.

A final substantive worry is a feature of the data for the consolidated complaints. In the late 1990s there was an alleged practice of using tie-in agreements as part of initial public offerings. These agreements allegedly required initial IPO investors to purchase additional shares in the aftermarket and, in some cases, at specified, increasing prices. A substantial number of lawsuits followed and these cases were consolidated into multi-district litigation ("MDL") before Judge Shira Scheindlin in the Southern District of New York ("SDNY"). There are 82 of these cases in the data set and they all resulted in a settlement. The consolidated complaints for these cases are highly similar and that creates the risk that the machine learning models will seek to identify these cases at the expense of identifying more general indications of quality. The inclusion of these cases may increase the estimated accuracy of the models because all of these sorts of cases will be classified correctly.

To assess this possibility, I repeat several of the tests with the SDNY MDL cases omitted. The average accuracy rate for 100 iterations of the Extra Trees classifier is 64.8% and the equivalent figure is 64.0% for the Random Forest classifier. These drops are modest; less than one percent for the Extra Trees classifier and slightly more than one percent for the Random Forest models. But eliminating 82 cases that settled drops the baseline settlement rate to 55.1% as opposed to a 56.7% settlement rate for the full sample of consolidated cases. The performance of the models without the MDL cases is thus a slight improvement over the models that use the full sample. This finding is consistent with the notion that eliminating the MDL cases allows the models to focus on more general indications of case quality rather a single type of case that produces many settlements.

4.3 Comparing Text-Only Models to Models that Incorporate Non-Text Features

One question of interest is how models that exclusively use text compare to those that incorporate the non-text variables that are often used in securities litigation research. It is possible that text-only analysis is able to capture these non-text variables and thus using non-text features may not do much to improve the models. Alternatively, non-text variables may capture case characteristics in a more accurate and parsimonious way. If so, the use of these variables may improve performance when used in combination with the text features or on their own (Geigle, Mei, and Zhai 2018).

This analysis poses some methodological challenges. Adding more variables to the 1500 features used in the bag-of-words analysis in the previous section produces degraded performance. That may be because the use of additional features produces overfitting in the model. To address this concern I reduce the dimensionality of the text through word embeddings. To do so, I use word2vec to create the word emdeddings, which, in contrast to the bag-of-words approach used above, allow for contextual analysis. The process maps the individual words in a corpus—in this case the universe of all first-filed and consolidated complaints—into vector space with 120 dimensions. The methodological appendix describes this technique in more detail. These 120 variables are the inputs for the text-only analysis. To perform the non-text analysis, I use the disclosure dollar loss and one-hot-encoded vectors for the court district, the year of filing, SIC industry of the defendant firm, and whether the complaint presents claims under Section 10(b) of the Securities Exchange Act of 1934 and/or Section 11 of the Securities Exchange Act of 1933. The combined analysis uses both all 120 text variables and and all of the non-text variables.

These methodological issues mean that the analysis that follows is not a comparison of the best performing text models and the best performing non-text models. None of the models used in this comparison, be they the further reduced text only models, non-text models, or the combined models, perform as well as the more complete text-only models presented in Table 2. The best absolute comparison between text-only models and non-text models is between the best performing text-only models in Table 2 and the best performing non-text models in Table 3. That comparison shows that the text models outperform the non-text models. While the advantage of the text models is not very large, it is strongly statistically significant.[8]

The goal of using modified text models and non-text models is to assess how these inputs perform independently and how well they predict when combined. This analysis, in addition to a comparison of the best performing text and non-text models, should provide some insight into the amount of information that may be lost when an analysis only uses non-text inputs. Table 3 reports the results of three different types of analysis

[8]To compare the best performing text model (the Random Forest model of first-filed cases from Table 2, Panel A) to the best performing non-text model (the AdaBoost model non-text model from Table 3, Panel A) I perform a two-sided Wilcoxon signed rank test on 100 accuracy rates from each model. This test is appropriate because the accuracy rates for each fold are not independent. The difference between the samples is significant at the .0001 level.

for the first-filed complaints (Panel A) and the consolidated complaints (Panel B). For the first-filed complaints, the text-only models perform better than those that only use non-text variables by a moderate amount. The models that combine both text and non-text features produce more accurate predictions than those that use only the text or only the non-text features. This increase is fairly substantial for the first-filed complaints. The results are a bit different for the consolidated complaints. The non-text models perform quite poorly relative to the text-only models and combining text and non-text features does not do much to improve performance if at all. For example, the random forest classifier is correct 64.6% of the time using only the text vectors, 61.2% of the time using non-text variables, and 64.4% of the time using the combined data.

Another way to depict the differences between the models is through receiver operating characteristic ("ROC") curves. These curves plot the true positive against the false positive rate as the threshold for discrimination varies. A perfect classifier would classify all cases exactly and would have complete confidence in all of those classifications. A classifier operating at random would, on average, fail to distinguish between true positives and false positives (i.e. would fall along the 45-degree line of the plot). A minimally competent classifier will show a curve that is above this 45-degree line. The area under this curve ("AUC") can indicate the strength of a classifier. A perfect classifier will have an AUC of 1 and a random classifier will have, on average, an AUC of .5.

Figure 2 depicts the ROC curve for the first-filed complaints for models that use only text features (i.e. the 120 document vectors), only the non-text variables, and both the text and non-text features. The curves show the interpolated results of 100 iterations that use the extra trees classifier. This figure is consistent with the results in Table 3. The text-only model outperforms the model that relies only on non-text variables (AUC of .74 versus .71). The combined model improves on the performance of the text-only model (AUC=.76). Figure 3 shows a different pattern for the consolidated complaints. The text model does much better than the non-text model (AUC of .71 versus .61) and roughly equivalent performance for the text-only and combined models (AUC both equal to .71).

These results, when combined with those that show that the best text models beat the best non-text models by modest margins, have two implications. First, text appears to do at least as well as non-text. Indeed, the strongest performing model across all the analyses is a text-only model, which suggests that text alone may be sufficient for some analyses. Second, the analysis in this subsection shows that combining text and non-text sources of information typically outperforms modified text-only models and non-text only models. These two implications sound a note of caution for securities litigation research that does not control for the textual content of complaints. Not taking account of textual variation may be a source omitted variable bias when using non-text variables in regression analysis. This bias may produce results that attribute effects to non-text variables, such as the statutory basis for the claim, the effect is actually driven by variation embedded in the text, such as the presence of hard facts in the case. This section also shows that it is not trivial to find a way to combine these sources of information in an efficient way. Future work in this area should consider both finding the best methods for combining textual and non-textual sources of information and using those methods to account for textual content.

4.4 Text Informativeness and Security Prices

A key question is how the machine learning models compare—both in the terms of speed and accuracy—to informed human judgment. Perhaps the best assessment of this sort would be to compare the predictions of the text-based models to those of experienced securities practitioners. Obtaining those types of predictions in numbers that provide sufficient statistical power would be prohibitively costly, but the securities markets provide another avenue of obtaining information about how informed market participants evaluate the information in a complaint. Previous research shows that stock market participants accurately incorporate expectations about securities litigation into stock prices within a few days of the complaint getting filed (Fich and Shivdasani 2007). The abnormal returns associated with text-based predictions thus serve as a potential means of external validation.

While the primary motive in conducting this analysis is to provide another means of assessing the informativeness of the text in the complaints, this exercise also contributes to the literature on complex information and asset pricing. The key question here is whether the predictions based on machine learning are quicker and more accurate than those of market participants. Cohen et al. (2020) attribute their finding that changes to the discussion of risk factors in annual reports and quarterly filings are associated with significant negative abnormal returns to both the complexity of the underlying language and investor inattention. They speculate that inattention may be what drives their result because the incorporation of the changes to the text into security prices can take weeks or even months. That inattention is probably less likely in the context of securities litigation, which in many cases is driven by newsworthy events such as alleged financial irregularities, accounting misstatements, and government investigations. But the information contained in the complaints can be complex and difficult to process. If that complexity introduces delay into assessing the likely consequences of a lawsuit, it is possible that machine learning models–which can produce an assessment more or less instantaneously–can generate abnormal returns if those models are able to evaluate cases with some accuracy. This type of result would be similar to that of Lee (2012), who finds that quarterly disclosures that are longer and more complex to read lengthen the stock price discovery process and prolong the post-earnings price drift.

To carry out this analysis, I obtain the $[t, t+10]$ defendant firm returns for the first-filed complaints from CRSP where the event ($t = 0$) is the filing of the complaint. I calculate the cumulative abnormal return using a market-adjusted model, CAPM, and Fama-French three and four-factor models. The market return for all models are estimated over the $t = -300$ to $t = -50$ window. Figure 4 shows the Fama-French four-factor cumulative abnormal return ("CAR") for the first-filed complaints that result in dismissal and those that result in settlements. The cases that will eventually settle have substantially lower abnormal returns than those that will get dismissed. This figures thus replicates the result from earlier work that shows that market participants are able to anticipate securities litigation outcomes when the complaint gets filed (Kempf and Spalt 2018; Griffin, Grundfest, and Perino 2004). It is worth noting that the returns are substantially negative on the day that the complaint gets filed. That is likely because some of these complaints get filed in the days after the market learns of the alleged fraud that is the basis of the securities complaint. The negative returns on the day of filing are likely to reflect the impact of that news in ways that reflect negative expectations about the firm, some of which are related to litigation and others that are not.

I next compare the returns based on actual to outcomes to those based on predicted outcomes. To make this assessment, I generate predictions using the Extra Trees model used to classify cases above. In the 100 iterations performed for each model, every case is in the held-out set an average of ten times. To produce a prediction for this analysis I average all the predictions for each case. Each individual prediction is the percentage of trees in the 165 trees used in each iteration of the model that predicts that the case will settle. As in Cohen et al. (2020), I focus on the cases in the highest quintile (i.e. those predicted to be most likely to settle) and the lowest quintile (i.e. those most likely to get dismissed) and then obtain the daily abnormal returns for the $[t, t+3]$, $[t, t+5]$, and $[t, t+10]$ periods. As discussed below, robustness checks show that the results in this analysis are not sensitive to this division.

Table 4 shows the returns for all four models on a portfolio that goes long on the cases in the lowest quintile and short on those in the highest quintile. I report the abnormal returns and the Sharpe ratios for each of these windows. It is clear that there are substantial, positive abnormal returns to the long-short portfolios. In all windows, in all four models, the returns are positive and statistically significant at the one-percent level. The returns are all above at least three percent and are nearly six percent in some cases. It appears that anyone trading on the textual information would be compensated for the risk, at least in first few days after filing. The Sharpe ratio is above 1.0 for all of the models in the $(t, t+3)$ window. This compares to the historical Sharpe ratio of roughly 1.0 for the S&P 500. As the trading window extends, returns appear to be more volatile, as all of the Sharpe ratios are below 1.0 for the $(t, t+10)$ windows. It thus appears, at least in the short term, that there is a delay to incorporating all the information in a complaint, even when accounting for the risk of trading on that information.[9] Figure 5 shows the abnormal returns in the ten days after filing for the lowest and highest quintiles for first-filed cases using the Fama-French four factor model. The returns for the cases that are predicted to be most likely to settle tracks pretty closely to the returns for cases that actually settle. Both show sharp drops on the day of filing. As discussed above, this is likely due to the market absorbing information about the underlying even t as well as the information about the lawsuit. Those cases predicted most likely to get dismissed have abnormal returns that are close to zero. While these returns are level throughout the $(t,\ t+10)$ period, the cases predicted most likely to settle do not level out until about four days after filing.

To assess the robustness of these results, I address two primary concerns. The first concern is that these results may be an artifact of the choice to a subset the predictions into the highest and lowest quintiles. Tables A3 and A4 in the Appendix assess the robustness of the results through similar analysis to that in Table 4, with one using the highest tercile and lowest terciles of predictions and the other using the entire dataset split at the median prediction. As in Table 4, there are positive abnormal returns to all the long-short portfoliios in all four models and the returns are all statistically significant

[9] I conduct a similar, unreported analysis for the consolidated complaints. If I include all of the consolidated complaints there are substantial and statstically signficiant abnormal returns in the ten-day window after the filing of these complaints. But these results appear to be an artifact of the 82 MDL cases discussed above. Nearly all of these cases were filed on the same day and the text of the complaints was virtually identical. Many of the defendant companies in these suits experienced declines in their stock price in the ten-day window following the filing of the complaints. It is unclear whether these declines were related to the litigation because most of these firms were technology businesses and this time period was a particularly volatile one for the equity prices of technology firms. When I omit these cases, the abnormal returns for all four models are near zero and none of them are statisically signficiant. This suggests that market participants have incorporated expectations about consolidated complaints into stock prices at the time of filing or that they are unable to do so within the ten-day post-filing period.

at the one-percent level. As one would expect when including a weaker set of predictions, the abnormal returns associated with the data split into the highest and lowest terciles are lower than when the data is split into the highest and lowest quintiles (the differences between the latter abnormal returns and the former abnormal returns range from .002 to .015). Likewise, the splitting the entire dataset at the median produces lower abnormal returns than subsetting the data into the highest and lower terciles (the differences between the latter abnormal returns and the former abnormal returns range from .013 to .019). These robustness tests help to confirm that the predictions based on text are associated with positive abnormal returns and they also provide evidence that the strength of the prediction is associated with the beliefs of stock market participants about the strength or weakness of a case.

The second concern is whether these types of abnormal returns could plausibly be obtained given their relatively large magnitude. This concern is, in some sense, collateral to the primary question of interest, which is whether the text alone can distinguish the good cases form the bad cases with the abnormal returns serving as a proxy for those good cases and bad cases. Nevertheless, it is important to note that for a significant number of the class actions it is unlikely that the returns in Table 4 could be obtained due to the timing of both the stock price effects and the filing of the complaints. Of the 3386 first-filed complaints, 790 of them were filed on the same date as the class-end date. That date is typically when the defendant firm makes a corrective disclosure to the market and that disclosure will often have a negative effect on the stock price.[10]

As Figure 4 shows, the decline on the day the complaint gets filed is sharper for the stronger cases than it is for the weaker cases (as measured by eventual outcome). Presumably, most of the complaints that get filed on the same day as the end of the class period get filed after the initial stock price reaction to the news. That means that any gain that would be had from shorting the cases that are predicted to be strongest would omit the initial drop. To assess the impact of this timing consideration, I calculate the abnormal returns for the $(t+1,\, t+3)$, $(t+1,\, t+5)$, and $(t+1,\, t+10)$ windows for the strongest and weakest quintiles of predictions for the first-filed cases. Using this strong assumption that no stock price movements could be captured on the day the complaint gets filed produces more modest, but still substantial, abnormal returns. For the returns on the long-short portfolios range from 2.1 to 2.5 percent for the $(t+1,\, t+3)$ window,

[10]Beyond this timing issue, there are additional considerations that would affect actual returns based on this strategy such as trading costs and borrowing costs associated with shorting. Moreover, if the stock of a sued company is relatively illiquid, trading could move prices in a way that further diminishes returns.

from 2.6 to 3.2 percent for the $(t+1,\ t+5)$ window , and from 3.0 percent to 4.1 percent for the $(t+1,\ t+10)$ window. All results are statistically significant at the one-percent level. The Sharpe ratios are around one for the $(t+1,\ t+3)$ windows, but are lower than one for the two longer windows. These results suggest that, even without trading on the day the complaint gets filed, there are substantial abnormal returns associated with trades based on text-based predictions.

These results bolster the conclusion that the text of securities complaints does include meaningful information about the strength or weakness of the case. Beyond showing an association with the beliefs of stock market participants, the analysis also shows that stronger predictions are associated with larger abnormal returns. This suggests that there are indications in the text that are able to distinguish some combination of the magnitude of the fraud and the magnitude of the damages. To put this another way, without providing any information about the alleged damages, the ultimate settlement, or how quickly the case was filed, analysis of the text alone provides indications about what happened—or what very recently happened—to the defendant's stock price. As the robustness checks show, there are reasons to doubt whether traders could capture the full extent of the abnormal returns in Table 4.

The robustness checks also show that even delayed trading based on the machine learning predictions produces substantial abnormal returns. This is because the predictions are instantaneous while the actual trading behavior appears to be subject to drift. Unlike complex earning information or changes to disclosures, the lawsuit drift lasts only a few days. As Figure 5 implies, the abnormal returns associated with the machine learning predictions level off after three or four days. Unreported results confirm this: the abnormal returns for the $(+5, +10)$ window based on a highest and lowest quintile analysis show abnormal returns of less than a percentage point and are statistically insignificant for all of the models except the market-adjusted model (which is only significant at the ten-percent level). The reason why the drift period is shorter for lawsuits than for complex earnings or disclosure changes is a question for future research, but some possibilities are that lawsuits are less complex than earnings or disclosure details or that the information underlying lawsuits is more salient than most information in a quarterly or annual filing.

5 Conclusion

This paper explores the degree to which text analysis and machine learning can predict outcomes in securities litigation. The strongest performing models can anticipate whether a case will settle or get dismissed with about 70 percent accuracy, which is a substantial improvement over baseline rates, although it leaves substantial room for improvement. A comparison of text-based models to non-text models shows that the text-based models modestly outperform the non-text models. When the text models are modified to allow a mixed text and non-text model, this combined approach outperforms models that only the modified text inputs or the non-text inputs. These findings suggest that models that use only non-text variables may be omitting information that can be distilled from the text. Future research should focus on techniques that can integrate information obtained from textual sources with traditional non-text variables used in studies of law and finance.

The machine learning predictions based on the text are also strongly associated with abnormal returns in the stock market in the days following the filing of the complaint. This finding helps to validate that the text is somewhat informative because previous research shows that stock market participants are able to predict the likely result in securities cases, albeit with a delay of several days. These results suggest that, like changes to disclosures and detailed descriptions of earnings, securities complaints are complex sources of information that take time for markets to digest.

More broadly, this paper shows the promise and perils of using text analysis and machine learning techniques in studies of law and finance, which often allow researchers to draw on rich sources of text. The promise lies in the results showing that these techniques can extract meaningful information about likely outcomes and about asset prices. The perils lie in the relative modesty of the results. Preparing text data sources is often a resource intensive task and the improvement over non-text models is slight. The magnitude of these improvements may increase as processing power increases and techniques improve, but in an important sense, the exercise here provides some measure of vindication for low dimension, non-text approaches to law and finance research.

References

Aletras, Nikolaos, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. "Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective." *PeerJ Computer Science* 2: e93.

Badawi, Adam B, and Daniel L Chen. 2017. "The Shareholder Wealth Effects of Delaware Litigation." *American Law and Economics Review* 19 (2): 287–326.

Badawi, Adam B, and David H Webber. 2015. "Does the Quality of Plaintiffs' Law Firm Matter in Deal Litigation." *J. Corp. L.* 41: 359.

Bebchuk, Lucian, Alma Cohen, and Allen Ferrell. 2008. "What Matters in Corporate Governance?" *Review of Financial Studies* 22 (2): 783–827.

Buehlmaier, Matthias MM, and Toni M Whited. 2018. "Are Financial Constraints Priced? Evidence from Textual Analysis." *The Review of Financial Studies* 31 (7): 2693–2728.

Cain, Matthew D, Jill E Fisch, Steven Davidoff Solomon, and Randall S Thomas. 2019. "Mootness Fees." *Vand. L. Rev.* 72: 1777.

Cain, Matthew D, and Steven Davidoff Solomon. 2014. "A Great Game: The Dynamics of State Competition and Litigation." *Iowa L. Rev.* 100: 465.

Catan, Emiliano M, and Marcel Kahan. 2016. "The Law and Finance of Antitakeover Statutes." *Stan. L. Rev.* 68: 629.

Cawley, Gavin C, and Nicola LC Talbot. 2010. "On over-Fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation." *The Journal of Machine Learning Research* 11: 2079–2107.

Chen, Tianqi, and Carlos Guestrin. 2016. "Xgboost: A Scalable Tree Boosting System." In *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–94. ACM.

Choi, Stephen J. 2006. "Do the Merits Matter Less After the Private Securities Litigation Reform Act?" *The Journal of Law, Economics, & Organization* 23 (3): 598–626.

Choi, Stephen J, Jill E Fisch, and Adam C Pritchard. 2005. "Do Institutions Matter-the Impact of the Lead Plaintiff Provision of the Private Securities Litigation Reform Act." *Wash. ULQ* 83: 869.

Choi, Stephen J, Karen K Nelson, and Adam C Pritchard. 2009. "The Screening Effect of the Private Securities Litigation Reform Act." *Journal of Empirical Legal Studies* 6 (1): 35–68.

Choi, Stephen J, and Adam C Pritchard. 2016. "SEC Investigations and Securities Class Actions: An Empirical Comparison." *Journal of Empirical Legal Studies* 13

(1): 27–49.

Cohen, Lauren, Christopher Malloy, and Quoc Nguyen. 2020. “Lazy Prices.” *The Journal of Finance* 75 (3): 1371–1415.

Cox, James D, and Randall S Thomas. 2006. “Does the Plaintiff Matter-an Empirical Analysis of Lead Plaintiffs in Securities Class Actions.” *Colum. L. Rev.* 106: 1587.

Edmans, Alex, Diego Garcia, and Øyvind Norli. 2007. “Sports Sentiment and Stock Returns.” *The Journal of Finance* 62 (4): 1967–98.

Fich, Eliezer M., and Anil Shivdasani. 2007. “Financial Fraud, Director Reputation, and Shareholder Wealth.” *Journal of Financial Economics* 86 (2): 306–36. https://doi.org/https://doi.org/10.1016/j.jfineco.2006.05.012.

Freund, Yoav, and Robert E Schapire. 1996. “Experiments with a New Boosting Algorithm.” In *Icml*, 96:148–56. Citeseer.

Garcia, Diego. 2013. “Sentiment During Recessions.” *The Journal of Finance* 68 (3): 1267–1300.

Geigle, Chase, Qiaozhu Mei, and C Zhai. 2018. “Feature Engineering for Text Data.” *Feature Engineering for Machine Learning and Data Analytics, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series*, 15–45.

Gompers, Paul, Joy Ishii, and Andrew Metrick. 2003. “Corporate Governance and Equity Prices.” *The Quarterly Journal of Economics* 118 (1): 107–55.

Gonçalves, Teresa, and Paulo Quaresma. 2005. “Is Linguistic Information Relevant for the Classification of Legal Texts?” In *Proceedings of the 10th International Conference on Artificial Intelligence and Law*, 168–76. ACM.

Griffin, Paul A, Joseph A Grundfest, and Michael A Perino. 2004. “Stock Price Response to News of Securities Fraud Litigation: An Analysis of Sequential and Conditional Information.” *Abacus* 40 (1): 21–48.

Hanley, Kathleen Weiss, and Gerard Hoberg. 2010. “The Information Content of IPO Prospectuses.” *Review of Financial Studies* 23 (7): 2821–64.

Hoberg, Gerard, and Gordon Phillips. 2016. “Text-Based Network Industries and Endogenous Product Differentiation.” *Journal of Political Economy* 124 (5): 1423–65.

Katz, Daniel Martin, Michael J Bommarito II, and Josh Blackman. 2017. "A General Approach for Predicting the Behavior of the Supreme Court of the United States." *PloS One* 12 (4): e0174698.

Kempf, Elisabeth, and Oliver Spalt. 2018. "Litigating Innovation: Evidence from Securities Class Action Lawsuits." *University of Chicago, Booth School of Business Working Paper.*

Krishnan, CNV, Steven Davidoff Solomon, and Randall S Thomas. 2015. "Who Are the Top Law Firms? Assessing the Value of Plaintiffs' Law Firms in Merger Litigation." *American Law and Economics Review* 18 (1): 122–54.

———. 2017. "The Impact on Shareholder Value of Top Defense Counsel in Mergers and Acquisitions Litigation." *Journal of Corporate Finance* 45: 480–95.

La Porta, Rafael, Florencio Lopez-de-Silanes, Andrei Shleifer, and Robert W Vishny. 1998. "Law and Finance." *Journal of Political Economy* 106 (6): 1113–55.

Le, Quoc, and Tomas Mikolov. 2014. "Distributed Representations of Sentences and Documents." In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, II-1188-II-1196. ICML'14. Beijing, China: JMLR.org. http://dl.acm.org/citation.cfm?id=3044805.3045025.

Lee, Yen-Jung. 2012. "The Effect of Quarterly Report Readability on Information Efficiency of Stock Prices." *Contemporary Accounting Research* 29 (4): 1137–70.

Nyarko, Julian. 2021. "Stickiness and Incomplete Contracts." *U. Chi. L. Rev.* 88: 1.

Palau, Raquel Mochales, and Marie-Francine Moens. 2009. "Argumentation Mining: The Detection, Classification and Structure of Arguments in Text." In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, 98–107. ACM.

Pritchard, Adam C, and Stephen P Ferris. 2001. "Stock Price Reactions to Securities Fraud Class Actions Under the Private Securities Litigation Reform Act." *Michigan Law and Economics Research Paper*, no. 01-009.

Rauterberg, Gabriel, and Eric Talley. 2017. "Contracting Out of the Fiduciary Duty of Loyalty: An Empirical Analysis of Corporate Opportunity Waivers." *Colum. L. Rev.* 117: 1075.

Spamann, Holger. 2009. "The 'Antidirector Rights Index' Revisited." *The Review of Financial Studies* 23 (2): 467–86.

Sulea, Octavia-Maria, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P Dinu, and Josef van Genabith. 2017. "Exploring the Use of Text Classification in the Legal Domain." *arXiv Preprint arXiv:1710.09306*.

Wainer, Jacques, and Gavin Cawley. 2018. "Nested Cross-Validation When Selecting Classifiers Is Overzealous for Most Practical Applications." *arXiv Preprint arXiv:1809.09446*.

Weiss, Elliott J. 2008. "The Lead Plaintiff Provisions of the PSLRA After a Decade, or Look What's Happened to My Baby." *Vand. L. Rev.* 61: 543.

Wongchaisuwat, Papis, Diego Klabjan, and John O McGinnis. 2017. "Predicting Litigation Likelihood and Time to Litigation for Patents." In *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law*, 257–60. ACM.

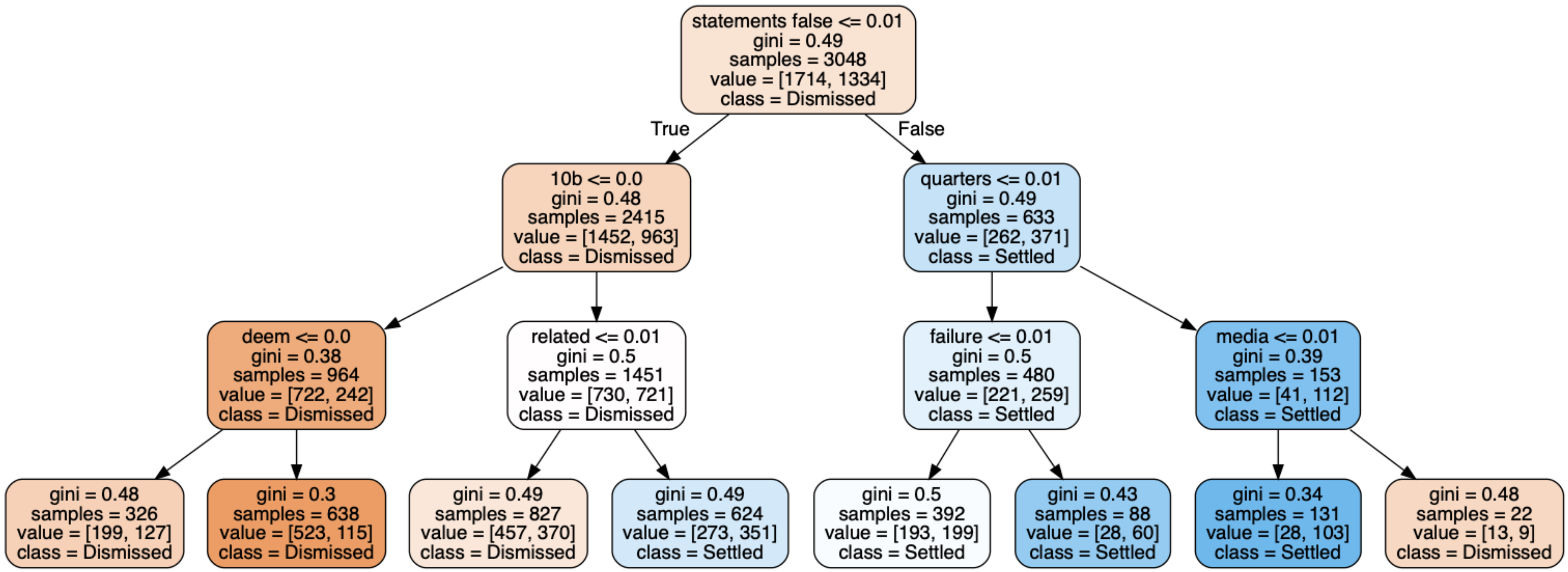

Figure 1: Example of a Simplified Decision Tree from a Random Forest Model

Note: The first line of each box indicates the term and threshold used to classify at that stage. The second line reports the Gini impurity for the term. The third line reports the number of cases that meet the criteria at that point of the decision tree. The fourth line reports the number of dismissed and settled cases at that point, respectively. The last line reports the classification—settled or dismissed—that is predominant at that point in the decision tree.

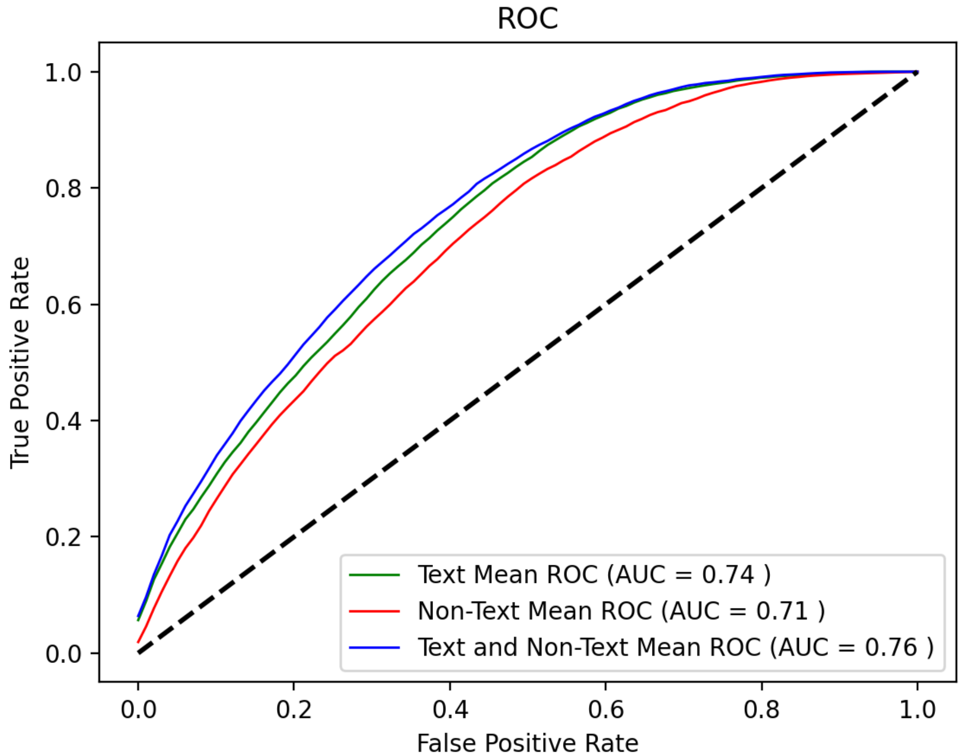

Figure 2: Receiver Operating Characteristic (ROC) Curve for First-Filed Complaints

Note: This figure reports the interpolated true positive rate against the false positive rate for different decision thresholds in 100 iterations of an Extra Trees classifier for the first-filed complaints.

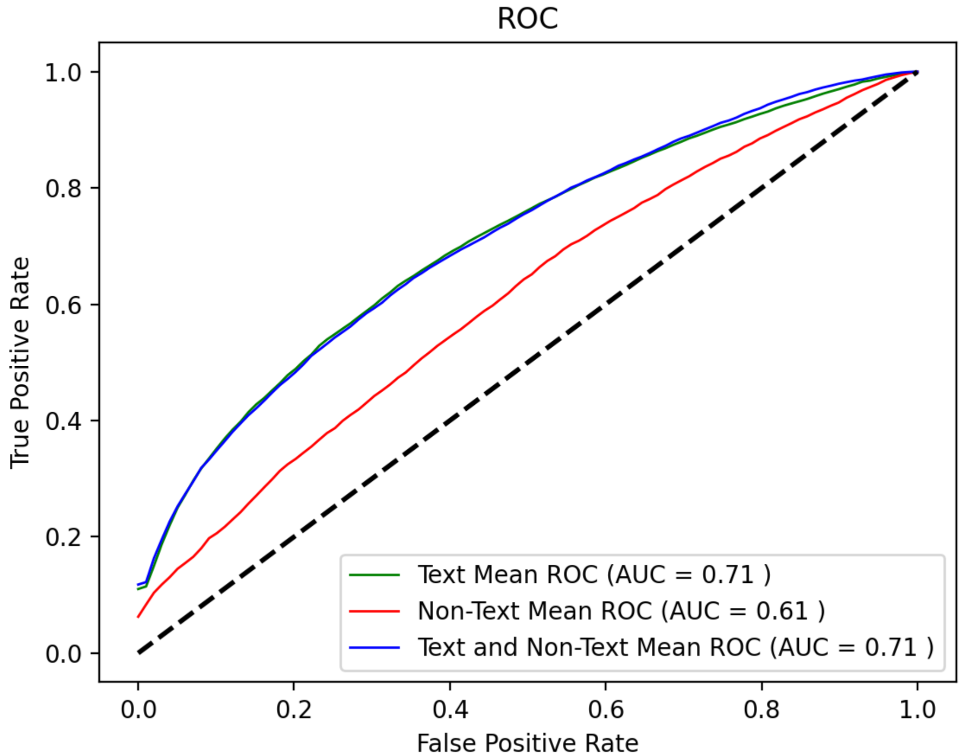

Figure 3: Receiver Operating Characteristic (ROC) Curve for Consolidated Complaints

Note: This figure reports the interpolated true positive rate against the false positive rate for different decision thresholds in 100 iterations of an Extra Trees classifier for the consolidated complaints.

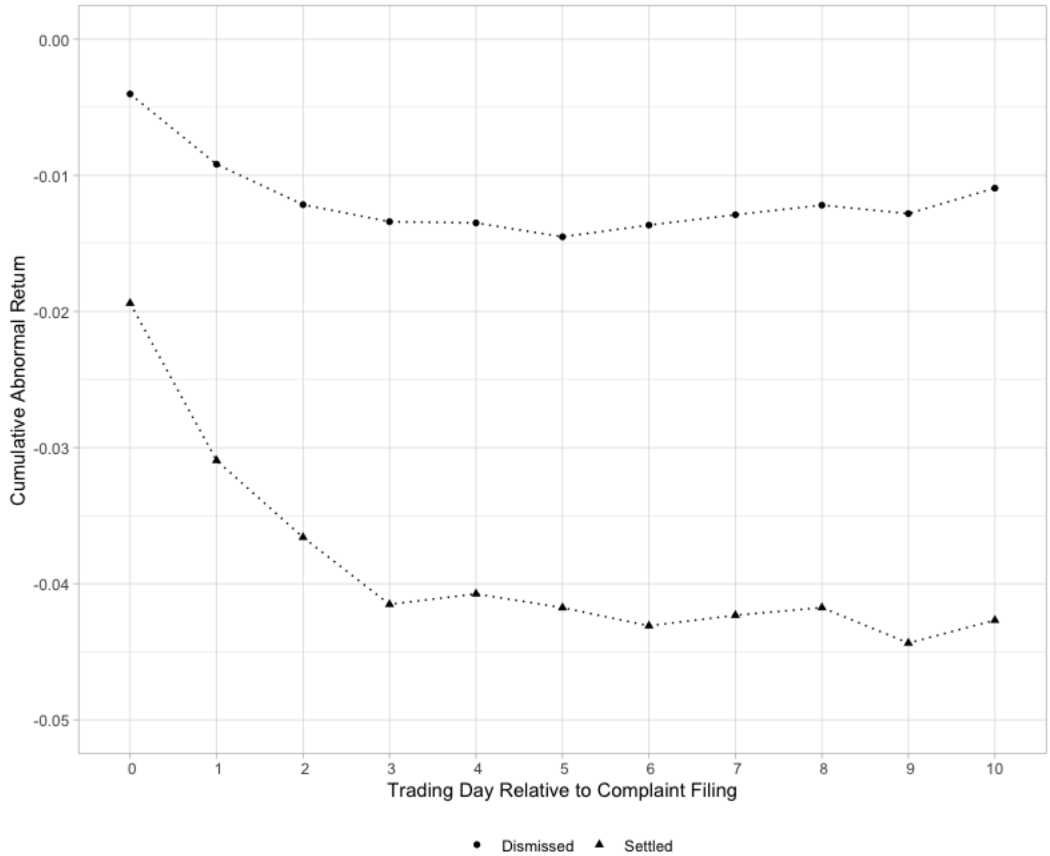

Figure 4: Cumulative Abnormal Return for First-Filed Cases Based on Eventual Outcome

Note: The figure shows the cumulative abnormal return over event days [0,+10] for first-filed cases based on whether those cases are actually settled or actually dismissed. The event is the filing of the complaint. The abnormal returns are those in excess of the four-factor Fama-French market return over days $t = -300$ to $t = -50$.

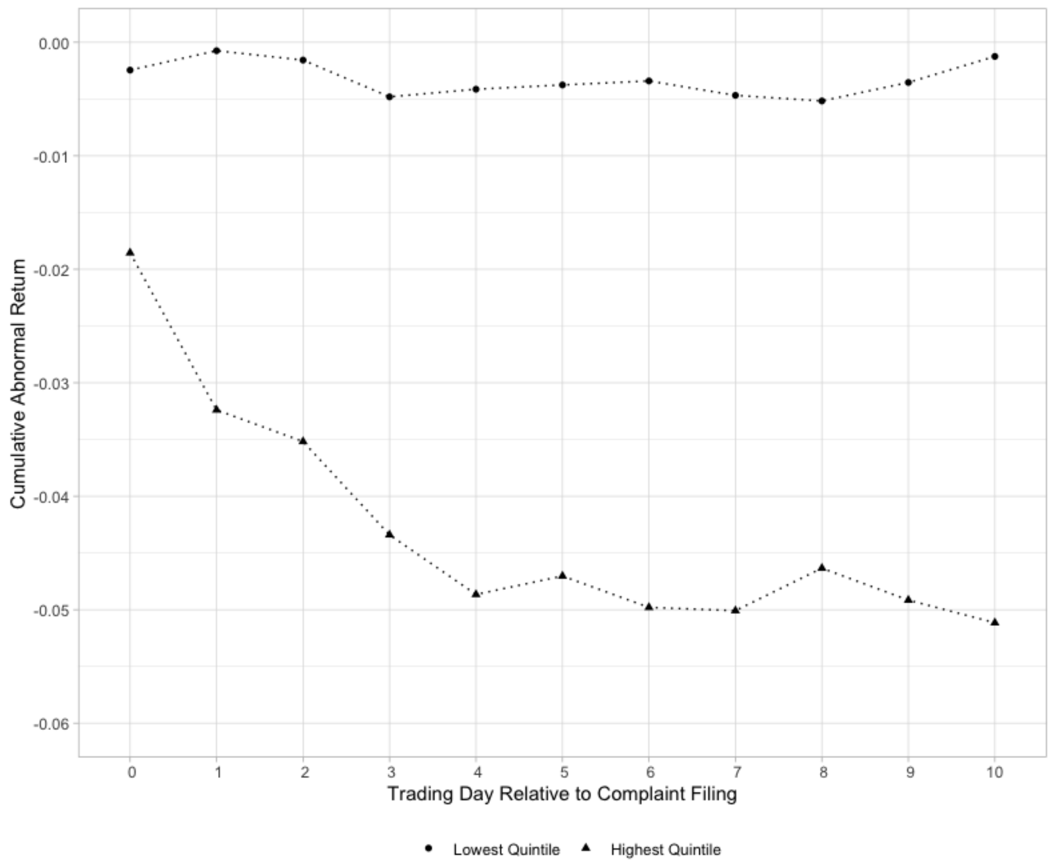

Figure 5: Cumulative Abnormal Return for First-Filed Cases in the Highest and Lowest Quintiles of Predicted Likelihood of Settlement

Note: The figure shows the cumulative abnormal return over event days [0,+10] for first-filed cases based on predicted likelihood of settlement. The predictions are derived from a random forest classifier that incorporates both text and non-text features. The abnormal returns are those in excess of the four-factor Fama-French market return over days $t = -300$ to $t = -50$.

Table 1: Summary Statistics for First-Filed Securities Cases (N=3386)

Statistic	Mean	Median	Min	Max
Year Initial Complaint Filed	2008	2008	1996	2019
Includes Section 10b Claim	0.769	1	0	1
Includes Section 11 Claim	0.136	0	0	1
Settled	0.437	0	0	1
Case Gets Consolidated	0.711	1	0	1
New York Court	0.231	0	0	1
California Court	0.212	0	0	1
Technology Industry	0.279	0	0	1
Words in Complaint	9150	7556	1492	161554
Words in Consolidated Complaints (N=2382)	24834	20041	2651	293160

Table 2: Classification Based on Textual Content of Securities Complaints

Panel A: First-Filed Case Predictions (N=3386, Baseline=56.3% Dismissed)

Classifier	Avg. Accuracy	2.5 Percentile	97.5 Percentile
Naive Bayes	.659	.619	.702
Random Forest	.699	.656	.739
ExtraTrees	.695	.655	.736
AdaBoost	.694	.649	.736
XGBoost	.678	.637	.717

Panel B: Consolidated Case Predictions (N=2382, Baseline=56.7% Settled)

Classifier	Avg. Accuracy	2.5 Percentile	97.5 Percentile
Naive Bayes	.617	.558	.690
Random Forest	.653	.607	.716
ExtraTrees	.654	.598	.703
AdaBoost	.650	.596	.711
XGBoost	.657	.603	.695

Table 3: Classification Based on Textual and Non-Textual Features of Securities Complaints Using Word Embeddings

Panel A: First-Filed Case Predictions (N=2702, Baseline=56.5% Dismissed)

Classifier	Avg. Accuracy	2.5 Percentile	97.5 Percentile
Text-Only Model			
Random Forest	.667	.627	.710
ExtraTrees	.667	.618	.727
AdaBoost	.650	.599	.696
XGBoost	.656	.607	.710
Non-Text Model			
Random Forest	.659	.605	.714
ExtraTrees	.639	.585	.697
AdaBoost	.664	.603	.712
XGBoost	.652	.598	.705
Combined Model			
Random Forest	.670	.603	.712
ExtraTrees	.685	.633	.747
AdaBoost	.657	.609	.708
XGBoost	.658	.613	.712

Panel B: Consolidated Case Predictions (N=1989, Baseline=56.3% Settled)

Classifier	Avg. Accuracy	2.5 Percentile	97.5 Percentile
Text-Only Model			
Random Forest	.646	.593	.696
ExtraTrees	.642	.580	.727
AdaBoost	.622	.550	.688
XGBoost	.634	.578	.681
Non-Text Model			
Random Forest	.612	.565	.658
ExtraTrees	.584	.528	.648
AdaBoost	.617	.553	.676
XGBoost	.592	.523	.648
Combined Model			
Random Forest	.644	.583	.701
ExtraTrees	.658	.583	.716
AdaBoost	.626	.560	.699
XGBoost	.630	.575	.686

Table 4: Long-Short Returns Based on First-Filed Case Predictions

	Market-Adjusted Model			CAPM		
	$(t, t+3)$	$(t, t+5)$	$(t, t+10)$	$(t, t+3)$	$(t, t+5)$	$(t, t+10)$
CAR	0.043***	0.050***	0.058***	0.037***	0.041***	0.045***
	(5.531)	(5.688)	(5.498)	(4.683)	(4.630)	(4.028)
Sharpe Ratio	1.551***	1.247***	0.848***	1.313***	1.016***	0.624***
	(5.396)	(5.556)	(5.39)	(4.567)	(4.529)	(3.964)
Observations	1010	958	922	1010	958	922
	Three-Factor			Four-Factor		
	$(t, t+3)$	$(t, t+5)$	$(t, t+10)$	$(t, t+3)$	$(t, t+5)$	$(t, t+10)$
CAR	0.035***	0.040***	0.045***	0.039***	0.043***	0.050***
	(4.443)	(4.365)	(4.023)	(4.834)	(4.703)	(4.403)
Sharpe Ratio	1.249***	0.961***	0.625***	1.360***	1.036***	0.685***
	(4.347)	(4.284)	(3.971)	(4.733)	(4.619)	(4.351)
Observations	1010	958	922	1010	958	922

Note: This table presents the cumulative abnormal return in the relevant time window for a portfolio that goes long on the cases in the lowest quintile of cases that are predicted to settle and short on the cases that are in the highest quintile. The relevant models used to obtain the abnormal returns are noted in the tables. I report the equal-weighted mean of the long-short portfolio with t-statistics in parentheses. The Sharpe ratio is the annualized value of the long-short returns divided by the standard deviation of those returns. Statistical significance is denoted by * p<.10, ** p<.05, *** p<.01.

Appendices

Appendix A: Technical Appendix

This technical appendix describes the processing of the text corpus and provides details on the machine learning techniques used in the analysis.

Text Processing.

Natural Language Processing ("NLP") analysis requires extracting and cleaning the raw text from the text, html, and pdf files obtained from SCAC. I use standard NLP protocols to do so. The text is converted to lower case and stop words—a list of the 174 most common words in the English language—are removed. Any words that do not appear in at least five of the documents are removed and the program then selects the 1500 most frequently used unigrams (single words) and bigrams (two-word combinations) in the corpus.

Much of the analysis uses term-frequency, inverse document frequency (tf-idf) transformation for as the primary input. This common technique converts a matrix of word counts into a matrix where the weights for each word increase as the word appears in a document and further increase if that word appears rarely in the entire corpus. This approach thus attaches more importance to words that are common in a document but uncommon across documents.

Machine Learning Models.

One of the oldest and most straightforward classification algorithms is the Naive Bayes family of algorithms. These approaches apply Bayes Theorem to generate predictions based on conditional probabilities associated with the different classes. What makes the approach naive is that it assumes independence among the features. Despite this unrealistic assumption, the model performs quite well for a number of tasks such as the identification of junk email. The model can be tuned by specifying the prior probabilities associated with each class.

Decision tree algorithms attempt to identify the model features that have the best ability to classify. A simple decision tree model that uses all available information is prone to overfitting and thus a method to improve prediction are so-called random forest models. These approaches select a random subset of features and generate a decision tree based on that subset. The researcher specifies the number trees that will be generated and each tree "votes" on the classification. The prediction is whatever category gets a majority of votes. The percentage of votes in favor of a classification can also serve as a measure of the strength of the random forest prediction. I use both the standard random forest model and the extra trees classifier. The difference between these models is that random forests choose the locally optimal feature split while extra trees split at random values.

One particularly successful classification strategy is to combine guesses from multiple algorithms. Random forests are an example of this type of boosting because they combine the results from many decision trees. Two of the other boosting approaches that I use, the Adaboost and XGBoost algorithms, use model stacking. Adaboost iteratively trains the algorithm by selecting the training set based on the accuracy of previous iteration. The final predictions reflect the weight of each classifier, which depends on the performance of that classifier (Freund and Schapire 1996). This technique uses multiple of the above methods to produce classifications and then uses a meta-classifier that relies on the output of the initial models to make predictions. This approach will often produce higher accuracy than those achieved with any single model. XGBoost is a type of gradient model, which constructs new models that seek to minimize the residuals from earlier models (Chen and Guestrin 2016).

Nested Cross-Validation.

To assess the robustness of the results in Table 2, I perform nested cross validation. In nested cross validation, the researcher specifies M folds for an outer loop, N folds for an inner loop, and possible parameters. For each outer loop, one m fold is held out and the M-1 folds are used to tune the hyperparameters in the inner loop. For the inner loop, one n fold is held out and the model is trained on the N-1 folds with each permutation of the hyperparameter settings. This process is repeated until each N folds has been held out and the performance metrics are calculated for each hyperparameter permutation. The best performing set of hyperparameters is then used to train a model on all N folds and that model is used to evaluate performance on the held out m fold. Performance is then averaged across all M folds.

In the robustness check, I specify 4 folds for both the outer loop and the inner loop. The set of parameters varies the number of features (1000, 1500, or 2000), the number of n-grams (only unigrams or both unigrams and bigrams), and the number of estimators used in the model (100, 150, or 200). I perform this process 25 times, which produces 100 estimates on the held-out portion of the outer fold. I focus on the Extra Trees and Random Forest classifiers as these models tend to perform the best across the analysis. Table A1 below presents the results of this exercise.

Table A1: Classification Using Nested Cross Validation

Panel A: First-Filed Case Predictions (N=3386, Baseline=56.3% Dismissed)

Classifier	Avg. Accuracy	2.5 Percentile	97.5 Percentile
Random Forest	.691	.670	.719
ExtraTrees	.690	.668	.715

Panel B: Consolidated Case Predictions (N=2382, Baseline=56.7% Settled)

Classifier	Avg. Accuracy	2.5 Percentile	97.5 Percentile
Random Forest	.650	.618	.680
ExtraTrees	.653	.617	.695

Word Embeddings.

To create word embeddings I use the word2vec algorithm on the entire corpus, which converts one-hot encodings of the cleaned complaints into a probability distribution that any word w will appear in proximity to word u. This process is done through a neural network that uses back propagation and gradient descent to produce the probability distributions. I use a model with skip gram architecture to produce the word embeddings across 120 dimensions (Le and Mikolov 2014). This process appears to work well as the five most proximate words to "stock" in the corpus are "price," "share," "trade," "common," and "sold" while the five closest words and word stems to the word stem "fals" are "materi," "made," "disclos," "knew,"" and "incomplet." To reduce each document down to a single 120-vector representation, I weight each word using tf-idf and then take the weighted average of the dimensions for each word that appears in a complaint.

Appendix B: Term Importance

The machine learning models can provide some indication of the terms that are most useful in discriminating between cases that get dismissed and those that settle. Table A2 provides the words that are most important for classification using the text-only random forest models for the first-filed and consolidated complaints. The random forest models are non-linear and thus it is not possible to associate words with a particular type of outcome. It may be that a certain word used along with other features is strongly predictive of dismissal while that same word in conjunction with other features is strongly predictive of settlement.

Table A2: Top 20 Most Important Terms for Classification in Random Forest Models

First Filed Complaints		Consolidated Complaints	
Term	Importance	Term	Importance
purchased	0.00588	investigation	0.00386
14a	0.00527	independent	0.00331
fails disclose	0.00522	staff	0.00325
prices	0.00453	signed	0.00321
material information	0.00445	improperly	0.00301
proposed	0.00444	registration statement	0.00266
artificially	0.00437	improper	0.00247
artificially inflated	0.00423	written	0.00241
cv	0.00381	required	0.00241
case	0.00363	accounting principles	0.00232
class period	0.00361	101	0.00205
reports	0.00333	contained	0.00204
fails	0.00311	plaintiff	0.00203
section 14	0.00311	financial statements	0.00201
suffered	0.00307	named	0.00199
fraud	0.00287	transactions	0.00198
inflated prices	0.0027	registration	0.00197
advisor	0.00268	principles	0.00196
inflated	0.00268	economic	0.00195
securities litigation	0.00267	practice	0.00193

Appendix C: Abnormal Return Robustness Checks

Table A3: Long-Short Returns Based on the Highest and Lowest Terciles of First-Filed Case Predictions

	Market-Adjusted Model			CAPM		
	$(t, t+3)$	$(t, t+5)$	$(t, t+10)$	$(t, t+3)$	$(t, t+5)$	$(t, t+10)$
CAR	0.038***	0.041***	0.043***	0.033***	0.033***	0.030***
	(6.302)	(6.018)	(5.193)	(5.320)	(4.630)	(4.759)
Sharpe Ratio	1.364***	1.016***	0.620***	1.152***	0.805***	0.418***
	(6.137)	(5.858)	(5.088)	(5.183)	(4.639)	(3.435)
Observations	1701	1695	1686	1701	1695	1686
	Three-Factor			Four-Factor		
	$(t, t+3)$	$(t, t+5)$	$(t, t+10)$	$(t, t+3)$	$(t, t+5)$	$(t, t+10)$
CAR	0.033***	0.034***	0.034***	0.036***	0.036***	0.038***
	(5.240)	(4.755)	(3.910)	(5.630)	(5.088)	(4.284)
Sharpe Ratio	1.137***	0.806***	0.469***	1.222***	0.863***	0.514***
	(5.115)	(4.645)	(3.848)	(5.500)	(4.976)	(4.224)
Observations	1701	1695	1686	1701	1695	1686

Note: This table presents the cumulative abnormal return in the relevant time window for a portfolio that goes long on the cases in the lowest tercile of cases that are predicted to settle and short on the cases that are in the highest tercile. The relevant models used to obtain the abnormal returns are noted in the tables. I report the equal-weighted mean of the long-short portfolio with t-statistics in parentheses. The Sharpe ratio is the annualized value of the long-short returns divided by the standard deviation of those returns. Statistical significance is denoted by * p<.10, ** p<.05, *** p<.01.

Table A4: Long-Short Returns Based on the First-Filed Case Predictions that are Above and Below the Median Prediction

	Market-Adjusted Model			CAPM		
	$(t, t+3)$	$(t, t+5)$	$(t, t+10)$	$(t, t+3)$	$(t, t+5)$	$(t, t+10)$
CAR	0.023***	0.022***	0.027***	0.019***	0.017***	0.017***
	(4.626)	(3.927)	(3.886)	(3.874)	(2.940)	(2.388)
Sharpe Ratio	0.814***	0.539***	0.378***	0.682***	0.404***	0.233***
	(6.137)	(5.858)	(5.088)	(5.183)	(4.639)	(3.435)
Observations	2533	2546	2532	2533	2546	2532
	Three-Factor			Four-Factor		
	$(t, t+3)$	$(t, t+5)$	$(t, t+10)$	$(t, t+3)$	$(t, t+5)$	$(t, t+10)$
CAR	0.019***	0.017***	0.019***	0.021***	0.019***	0.023***
	(3.742)	(2.943)	(2.684)	(4.164)	(3.239)	(3.158)
Sharpe Ratio	0.659***	0.405***	0.262***	0.734***	0.446***	0.309***
	(3.634)	(2.859)	(2.635)	(4.407)	(3.151)	(3.106)
Observations	2533	2546	2532	2533	2546	2532

Note: This table presents the cumulative abnormal return in the relevant time window for a portfolio that goes long on the cases in the lower half of cases that are predicted to settle and short on the cases that are in the higher half. The relevant models used to obtain the abnormal returns are noted in the tables. I report the equal-weighted mean of the long-short portfolio with t-statistics in parentheses. The Sharpe ratio is the annualized value of the long-short returns divided by the standard deviation of those returns. Statistical significance is denoted by * p<.10, ** p<.05, *** p<.01.