How Informative Is the Text of Securities Complaints?

Adam B. Badawi*

September 30, 2019

Preliminary Draft: please do not circulate without permission.

Abstract

Using the complaints from several thousand securities class actions filed from 1996 to 2019, this paper uses text analysis and machine learning to predict whether those lawsuits will settle or be dismissed. The strongest performing models are able to predict the outcome in these cases at a rate of just over 70 percent, which is a substantial improvement over baseline settlement rates. The models are also able to generate a probability that each case will settle and these estimates provide strong indications of future equity returns. A portfolio that goes long on the firms with consolidated cases that are most likely to be dismissed and short on those consolidated cases that are most likely to settle produces abnormal returns of over three percent in the ten-day window that follows the filing of the complaint. Beyond contributing to the asset pricing literature, the findings have implications for several other areas of research. That it is easier to predict the outcomes of first-filed complaints relative to consolidated complaints provides empirical support for the notion that there is still something of a race to the courthouse in securities litigation. In addition, the predictive ability of the text in complaints suggests that variables built on these measures may help to control for case quality in studies of business litigation. Finally, while these models perform reasonably well, there is substantial room for improvement. This observation implies that, at least for the time being, predictive analytics should act as a complement to, rather than substitute for, human legal judgment.

[*] Professor of Law, UC Berkeley, Visiting Scholar, University of Amsterdam, Center for Law and Economics; abadawi@berkeley.edu. I thank Joe Grundfest for providing access to data from the Securities Class Action Clearinghouse. I also thank Bobby Bartlett, Steve Choi, Jonah Gelbach, Craig Lewis, Adair Morse, and participants at the 2019 Vanderbilt Law and Business Conference, and the Berkeley Center for Law and Business retreat for helpful comments.

1 Introduction

Predicting and explaining litigation outcomes is an important task for lawyers, their clients, and academic observers. While commentators have long thought that computation and automation would ease this endeavor, progress has been halting. Most studies that utilize machine learning to analyze text associated with cases do so based on materials produced after a court has rendered a decision (e.g., using the text of an opinion to predict whether that opinion rules in favor of a plaintiff or a defendant). While a handful of papers have attempted to predict either the instance or outcome of litigation using information that is available prior to a legal decision, none of these studies have used the text of litigation documents to assess the likely outcome of a case. This study attempts to do so by answering what is arguably one of the most useful predictive questions in litigation: can machine learning use the text of a complaint, which is usually the first document filed by a plaintiff, to predict whether a case is likely to settle or be dismissed? I answer this question in the context of securities fraud class actions, which are frequent, high stakes, and often produce substantial settlements. The most accurate models are able to correctly predict whether a case settles or gets dismissed just over 70 percent of the time. These findings have implications for several areas of law and finance.

First, the results contribute to the asset pricing literature by showing that markets are slow to react to the information contained in the text of securities complaints. The strongest predictions from the machine learning models developed here are associated with substantial abnormal returns over the ten-day window that follows the filing of a complaint. To take one example, a portfolio of defendant firms that that goes long on the consolidated cases that are predicted to be in the lowest quintile of likely settlement and short on the consolidated cases in the highest quintile of likely settlement would, on average, earn over three percentage points of abnormal return over the ten days following the filing of the complaint. These results provide another example in the literature of circumstances where text analysis can uncover pricing anomalies in financial markets (Cohen et al., 2018; Edmans et al., 2007; Garcia, 2013).

Second, this analysis helps 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).

Third, the analysis makes some progress on the difficult problem of controlling for the merits or strength of a lawsuit. 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, the nature of the fraud allegations, whether the SEC investigated the allegations, and the statute underlying the claim. Studies of merger class actions will typically control for the transaction size, party industries, transaction structure, and sometimes the law firms involved in the case. This paper shows that there is information about the likelihood of success contained in the text of the complaints in securities cases. Variables derived from this text, such as the predicted probability of producing a settlement, could plausibly be used to complement the other variables that are used to proxy for lawsuit quality.

Finally, this paper contributes to the literature on automation and legal analysis. There has been substantial debate and speculation about the effect that machine learning and other algorithmic analysis techniques will have on the work of lawyers (McGinnis and Pearce, 2013; Nissan, 2017). This relatively basic classification exercise shows that text analysis and machine learning have some capacity to predict outcomes in high stakes litigation. But the success rates are not all that high, which suggests that there is substantial room for machine learning and artificial intelligence techniques to improve in legal applications. In addition, the black box nature of machine learning provides relatively little information about the underlying basis for a prediction, which may also limit the ability of algorithms to replace human judgment and reasoning. Rather than suggesting that the robots are soon to take over, this study provides evidence that the best current use for machine learning and prediction is as a complement to the assessments of lawyers rather than as a substitute for them.

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 the data collection and the construction of the text corpus. Section 4, provides an overview of machine learning 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 and an Appendix provides information about the terms that are most useful in predicting lawsuit outcomes.

2 Literature Review

The 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 requires plaintiffs to plead "with 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 (i.e. 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 little difference 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 et al. (2009)

[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).

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 et al. (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, but I can 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 (-3.47 percent). Griffin et al. (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 have a more negative reaction to cases that alleged accounting issues. 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. Those choices should reflect something about the abilities of the plaintiffs' lawyers who draft the complaint and the underlying merits of the case. 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 et al. (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 et al., 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 business law 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 focused on identifying the topics of legal text (Gonçalves and Quaresma, 2005) and classifying argumentative versus non-argumentative text (Palau and Moens, 2009). More recent scholarship seeks to predict 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. But both of these studies use ex post measures; the text that they analyze is produced to justify the decision rendered.

There are fewer projects that use ex ante information to predict outcomes and, as far as I can tell, no other paper uses ex ante text produced by an adversarial party to predict litigation outcomes. Examples of these ex ante approaches include Katz et al. (2017), which uses non-text features to predict US Supreme Court outcomes with 70 percent accuracy and Wongchaisuwat et al. (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.

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. (2018) 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 goes long on those that change it the least earns up to 188 basis points in monthly alphas. 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 Securities Class Action Clearinghouse (SCAC). This database provides a comprehensive overview of securities filings in the United States going back to 1996. At present, SCAC contains information about 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 2018. 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.

Natural Language Processing ("NLP") analysis requires extracting and cleaning the text from the complaints. After extracting the raw text from the text, html, and pdf files obtained from SCAC, I use standard NLP protocols for cleaning the text. 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.[2]

Much of the analysis that follows uses term-frequency, inverse document frequency (tf-idf) transformation for the analysis. 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.

The machine learning analysis applies an algorithm to the weights from the tf-idf matrix to predict whether a securities case is dismissed or settled. As is standard, the success of this process is assessed using k-fold validation. This procedure starts by breaking the data into k partitions. I use k-1 partitions to build a predictive model of settlement

[2] The primary goal of limiting the number of features is to prevent overfitting models to the data.

and dismissal and then predict the held out partition with that model. This process is repeated until there is a prediction for every document in the data set. This exercise provides some sense of the ability of the model to predict outcome out-of-sample. 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 analysis, I use 10 folds to generate predictions.

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 a 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. This information is not available for all of the cases in the data sets. After merging together these data sets, there are 3270 first-filed cases and 2380 consolidated cases. Some of the additional analysis relies on security prices, which I obtain from the Center for Research on Security Pricing (CRSP).

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.[3] 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.

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 algorithms that I use and then

[3] 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.

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.[4] 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 initial and consolidated complaints.

4.1 Predicting Case Outcomes on the Basis of Complaint Text

The process for outcome prediction follows a typical machine learning classification analysis. 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.[5] 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.

One of the oldest and most straightforward classification algorithms is the Naive Bayes family of algorithms. These approaches apply Baye's 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. To take an example, one classic exercise in machine learning is to predict the survival of passengers on the Titanic based on passenger attributes. Gender, age, and title turn out to be very strong predictors of survival and additional information about, say, young aristocraatic women turns out not to have much predictive power. But for other groups of passengers, such as young men, additional information like class of ticket class or point of embarkation can help to refine predictions. A simple decision tree model

[4] 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.

[5] 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.

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.

Support vector machines are another popular technique for classification. This approach seeks to draw a hyperplane that splits the data in a way that minimizes errors. An advantage of SVMs is that they can perform linear transformations on non-linear data and then calculate the hyperplane that best splits the transformed data. For this reason, this technique, called the kernel trick, can be effective when there are non-linear relationships between the observations. SVMs can be tuned using a regularization parameter, which trades off accuracy against overfitting, and a gamma parameter which places different weights on outlying observations.

One particularly successful classification strategy is to combine guesses from multiple algorithms. The two 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 et al., 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).

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, Panel B provides the results for the consolidated complaint in cases that reach that stage, and Panel C provides the outcomes for combining all of the complaints into a single corpus. 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 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 number of correct and incorrect predictions and an accuracy score for each algorithm. In Panel A, the Random Forest and AdaBoost classifiers perform best. Both of these algorithms correctly classify whether a securities complaint is dismissed or settled around 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 support vector machine algorithm performs worst, but still provides the correct result in over 63% of the cases.

The results are somewhat different for the consolidated complaints in Panel B. The top classifiers for these complaints are the XGBoost and AdaBoost classifiers, which correctly classify outcomes 66.5% and 65.6% of the time respectively. The random forest classifier performs about a percentage point worse than the AdaBoost classifier. The support vector machine and Naive Bayes algorithms perform the worst on this task.[6] These results are against a baseline of 56.7% of the cases settling and are obtained 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.

The strongest improvements over the baseline are obtained by combining all of the complaints into a single corpus as Panel C shows. The random forest and AdaBoost classifiers produce the best results, classifying the outcome correctly 70.7 percent of the time. The improvement in performance may be due to an algorithm learning, in some cases, from two documents–the first-filed and consolidated complaints–that produced the same outcome. But the none of the algorithms are expressly informed that multiple documents are associated with the same case.

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

[6] 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.

likely to succeed than others. The terms that are most helpful in discriminating between outcomes, which are provided in Table A1 of the Appendix, provide some indication that the timing and topic of the case are useful in making predictions.

Second, 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 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. This evidence cannot make absolute quality comparisons between first-filed and consolidated complaints and it cannot make any comparisons between pre and post PSLRA cases.

Third, the ensemble methods consistently perform better than the more straightforward algorithms, such as Naive Bayes. This outcome is not that surprising given that ensemble methods aggregate information from across different algorithms. Nevertheless, the relatively poor performance of Naive Bayes suggests that the problem of classifying language in a complaint presents a complex problem that involves interactions between features and may have nonlinear characteristics.

4.2 Incorporating Non-Text Features

One way to improve classification is to combine text and non-text features (Geigle et al., 2018). This approach seems promising in the context of securities litigation where variables such as the court district or industry may be important predictors of litigation success or failure. Combining features in a way that improves classification typically requires reducing the dimensionality of the text features. A common way of doing so is

[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.

to perform truncated singular value decomposition on the tf-idf matrix (this process is also called latent semantic analysis). For both the first-filed and consolidated complaints the optimal number of dimensions is about 30. For the non-text features, the model uses one-hot-encoded vectors for the court district, the year of filing, SIC industry, 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.

Using combined methods makes it difficult or impossible to use some of the classification algorithms. For this reason, the analysis in Table 3 uses only the random forest and AdaBoost algorithms employed in the earlier results. The results show that incorporating the non-text features does not improve performance for any of the classifiers for both the first-filed and consolidated complaints. Taking into account the lower dismissal baseline of the cases used for the combined analysis, both algorithms perform about one and a half percentage points lower for the first-filed complaints and about two percentage points lower for the consolidated complaints. These results suggest two things about this classification problem. First, the text-only analysis may be picking up the non-text features used in this analysis–as well as others–that make the express inclusion of these non-text features redundant. Second, and relatedly, condensing the text down to just thirty dimensions may lose information that is more important than that included in the non-text features. Some combination of these two effects could account for the weaker performance that occurs when including non-text features.

4.3 Text Analysis and Security Prices

A natural comparator for prediction through text analysis is the stock market. Securities litigation can have meaningful financial consequences for firms (Fich and Shivdasani, 2007). Expectations about those consequences should be reflected in the trading of the stock of those firms. This subsection compares predictions inferred from stock prices to those obtained through the machine learning predictions. This subsection also assesses whether the predictions of case outcomes are associated with abnormal returns in equity markets.

I obtain the [0,+10] returns for all the first-filed and consolidated complaints from CRSP where the event ($t = 0$) is the filing of the complaint. I calculate the cumulative abnormal return using the Fama-French three factor model estimated over the $t = -300$ to $t = -50$ window. Figure 1 shows the cumulative abnormal return ("CAR") for the first-filed complaints that result in dismissal and those that result in settlements. As the figure shows, there is a substantial difference between the returns associated with cases that will eventually get dismissed (N=1700) and those that will eventually settle (N=1446). The mean CAR for dismissed cases is -.018 and -.048 for settled cases. That difference is statistically significant at the one-percent level (t-stat=4.68). This result is consistent with earlier work that shows that market participants are able to anticipate securities litigation outcomes when the complaint gets filed (Griffin et al., 2004; Kempf and Spalt, 2018). 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 (i.e. during the class period). 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.

Figure 2 shows the [0,+10] returns for the filing of consolidated complaints. Figure 2 differs from Figure 1 in several important respects. First, the CARs around filing are close to zero. This is what one would expect given that the underlying behavior is already known to the market. Second, market participants do appear to absorb the content of the complaint, but it takes time to do so. The most stark divergence does not appear until about five days after the filing of the consolidated complaint. Third, the ten-day CARs to cases that will be dismissed is positive and those that will be settled are negative. The difference in average ten-day CARs for dismissed cases (.010, n=898) and settled cases (-.004, n=1147) is statistically significant at the five-percent level (t-stat=2.10).

Another natural question is whether the predictions generated by text analysis are associated with abnormal returns in the stock market. To assess this question, I generate predictions using the random forest model with text and non-text features for the first-filed cases and the text features for the consolidated cases. The random forest method works by choosing a random set of features and building a decision tree based on those features. That decision tree is used to generate a prediction for every case and the overall prediction is based on the votes produced by each of these decision trees. A more precise probability is the overall percentage of the votes in favor of one classification. In this case, I use 150 trees and the predicted probability of settlement is the proportion of those 150 trees that classify the case as one that will settle. From these probabilities, I select 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 [0,+10] period.

Figure 3 shows the abnormal returns for the lowest and highest quintiles for first-filed cases. When compared to Figure 1, it shows that these predictions are associated with abnormal returns. Those cases that are predicted to be most likely to be dismissed have higher–although still negative–abnormal returns relative to cases that are actually dismissed. The difference on the tenth day is about one percentage point in abnormal return. The same is true for those cases that are predicted to be most likely to be settled. Those abnormal returns are more negative than the average abnormal return for the cases that actually settle (Figure 2). And again, the difference is about a percentage point. The difference in the ten-day mean CARs for the highest quintile (N=558) and lowest quintile (N=522) is about .044 and that difference is significant at the one-percent level (t-stat=4.39).

Figure 4 shows the abnormal returns for consolidated cases in the lowest and highest quintiles of likelihood to settle. For the lowest quintile, the CARs are not that different from the abnormal returns associated with actual dismissals. By the tenth day, the returns from the lowest quintile are lower than the actual dismissals although the difference is small. For those cases that are predicted to be most likely to settle, the difference with the cases that actually settle are large. By the tenth day, there is a difference of nearly four percentage points of abnormal return. The difference in the ten-day mean CARs for the highest quintile (N=272) and lowest quintile (N=371) is about .033 and that difference is significant at the five-percent level (t-stat=2.54).This result suggests that some consolidated complaints reveal important and damaging information about firms and that text analysis can identify those complaints.

These findings suggest that securities complaints contain important information about future returns for firms. This observation makes two contributions to the asset pricing literature. First, the evidence shows that there is a delay in incorporating information that machine learning models can identify quickly. This finding suggests that applying these sorts of computational techniques to unstructured text is likely to increase reaction times to market-relevant information. The second, and related, contribution is that the time to incorporate the content of securities complaints appears to be related to the complexity of that information. The abnormal returns for first-filed complaints in Figure 1 stabilize about three days after the filing of the complaint. For the consolidated complaints, which are, on average, three times longer than the first-filed complaints, prices do not level out until six to seven days after filing. This evidence is corroborated by the abnormal returns associated with the strongest predictions of the machine learning models. The first-filed cases that are predicted to be most likely to be dismissed–which are presumably among the lowest quality cases–hardly have any effect on the stock price, as Figure 3 shows. For those cases that are most likely to settle–and likely involve meritorious allegations that may be complex to understand–the stock prices do not level out until three days after filing. The predictions about consolidated cases suggest an even more dramatic story. In Figure 4, the abnormal returns for cases predicted to be likely to be dismissed do not move that much. But for cases that are predicted to settle, the abnormal returns steadily decline over the entire ten-day window. This delay in the full incorporation of the information for the most meritorious cases suggests that this information is complex and difficult to process. This result is consistent with studies finding that complex information takes a longer time to get absorbed into stock prices (Cohen and Lou, 2012; Cohen et al., 2018; Lee, 2012).

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. Generating more precise predictions from the machine learning models provides insight into whether the strongest predictions anticipate asset prices. This analysis shows that these predictions, which can be generated shortly after filing, are associated with up to five points of abnormal return in the ten-day window following filing. This finding contributes to the asset pricing literature on the absorption of complex information into security prices. Consistent with other work, the results in this paper suggest that more difficult cases take a longer time to process.

The evidence presented in this paper also contributes to studies of litigation and lawyering. The result that it is easier to predict the outcome of first-filed cases than consolidated cases corroborates accounts that some plaintiffs' attorneys still file hastily prepared complaints. The fact that text analysis of these complaints is moderately successful as a predictor shows that the words lawyers choose–either due to their own skill or due to the underlying facts of the cases they choose to bring–reflect the underlying quality of those cases. More broadly, the ability to predict outcomes with reasonable success solely on the basis of the text of the complaint shows both the promise and limitations of legal analytics. The promise suggests that machine learning can be useful as a way to analyze legal documents for quick assessments of merit. But these judgments leave some room for improvement in accuracy and a lot of room for improvement in their ability to explain why some cases are likely to succeed and others are not.

References

Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., Lampos, V., 2016. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. PeerJ Computer Science 2, e93.

Badawi, A.B., Chen, D.L., 2017. The shareholder wealth effects of delaware litigation. American Law and Economics Review 19, 287–326.

Badawi, A.B., Webber, D.H., 2015. Does the quality of plaintiffs' law firm matter in deal litigation. J. Corp. L. 41, 359.

Buehlmaier, M.M., Whited, T.M., 2018. Are financial constraints priced? Evidence from textual analysis. The Review of Financial Studies 31, 2693–2728.

Cain, M.D., Solomon, S.D., 2014. A great game: The dynamics of state competition and litigation. Iowa L. Rev. 100, 465.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. ACM, pp. 785–794.

Choi, S.J., 2006. Do the merits matter less after the private securities litigation reform act? The Journal of Law, Economics, & Organization 23, 598–626.

Choi, S.J., Fisch, J.E., Pritchard, A.C., 2005. Do institutions matter-the impact of the lead plaintiff provision of the private securities litigation reform act. Wash. ULQ 83, 869.

Choi, S.J., Nelson, K.K., Pritchard, A.C., 2009. The screening effect of the private securities litigation reform act. Journal of Empirical Legal Studies 6, 35–68.

Choi, S.J., Pritchard, A.C., 2016. SEC investigations and securities class actions: An empirical comparison. Journal of Empirical Legal Studies 13, 27–49.

Cohen, L., Lou, D., 2012. Complicated firms. Journal of financial economics 104, 383–400.

Cohen, L., Malloy, C., Nguyen, Q., 2018. Lazy prices. National Bureau of Economic Research.

Cox, J.D., Thomas, R.S., 2006. Does the plaintiff matter-an empirical analysis of lead plaintiffs in securities class actions. Colum. L. Rev. 106, 1587.

Edmans, A., Garcia, D., Norli, Ø., 2007. Sports sentiment and stock returns. The Journal of Finance 62, 1967–1998.

Fich, E.M., Shivdasani, A., 2007. Financial fraud, director reputation, and shareholder wealth. Journal of Financial Economics 86, 306–336. https://doi.org/https://doi.org/10.1016/j.jfineco.2006.05.012

Freund, Y., Schapire, R.E., others, 1996. Experiments with a new boosting algorithm, in: Icml. Citeseer, pp. 148–156.

Garcia, D., 2013. Sentiment during recessions. The Journal of Finance 68, 1267–1300.

Geigle, C., Mei, Q., Zhai, C., 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.

Gonçalves, T., Quaresma, P., 2005. Is linguistic information relevant for the classification of legal texts?, in: Proceedings of the 10th International Conference on Artificial Intelligence and Law. ACM, pp. 168–176.

Griffin, P.A., Grundfest, J.A., Perino, M.A., 2004. Stock price response to news of securities fraud litigation: An analysis of sequential and conditional information. Abacus 40, 21–48.

Hanley, K.W., Hoberg, G., 2010. The information content of ipo prospectuses. Review of Financial Studies 23, 2821–2864.

Hoberg, G., Phillips, G., 2016. Text-based network industries and endogenous product differentiation. Journal of Political Economy 124, 1423–1465.

Katz, D.M., Bommarito II, M.J., Blackman, J., 2017. A general approach for predicting the behavior of the supreme court of the united states. PloS one 12, e0174698.

Kempf, E., Spalt, O., 2018. Litigating innovation: Evidence from securities class action lawsuits. University of Chicago, Booth School of Business Working Paper.

Krishnan, C., Solomon, S.D., Thomas, R.S., 2017. The impact on shareholder value of top defense counsel in mergers and acquisitions litigation. Journal of Corporate Finance 45, 480–495.

Krishnan, C., Solomon, S.D., Thomas, R.S., 2015. Who are the top law firms? Assessing the value of plaintiffs' law firms in merger litigation. American Law and Economics

Review 18, 122–154.

Lee, Y.-J., 2012. The effect of quarterly report readability on information efficiency of stock prices. Contemporary Accounting Research 29, 1137–1170.

McGinnis, J.O., Pearce, R.G., 2013. The great disruption: How machine intelligence will transform the role of lawyers in the delivery of legal services. Fordham L. Rev. 82, 3041.

Nissan, E., 2017. Digital technologies and artificial intelligence's present and foreseeable impact on lawyering, judging, policing and law enforcement. Ai & Society 32, 441–464.

Palau, R.M., Moens, M.-F., 2009. Argumentation mining: The detection, classification and structure of arguments in text, in: Proceedings of the 12th International Conference on Artificial Intelligence and Law. ACM, pp. 98–107.

Pritchard, A.C., Ferris, S.P., 2001. Stock price reactions to securities fraud class actions under the private securities litigation reform act. Michigan Law and Economics Research Paper.

Sulea, O.-M., Zampieri, M., Malmasi, S., Vela, M., Dinu, L.P., Genabith, J. van, 2017. Exploring the use of text classification in the legal domain. arXiv preprint arXiv:1710.09306.

Weiss, E.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, P., Klabjan, D., McGinnis, J.O., 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. ACM, pp. 257–260.

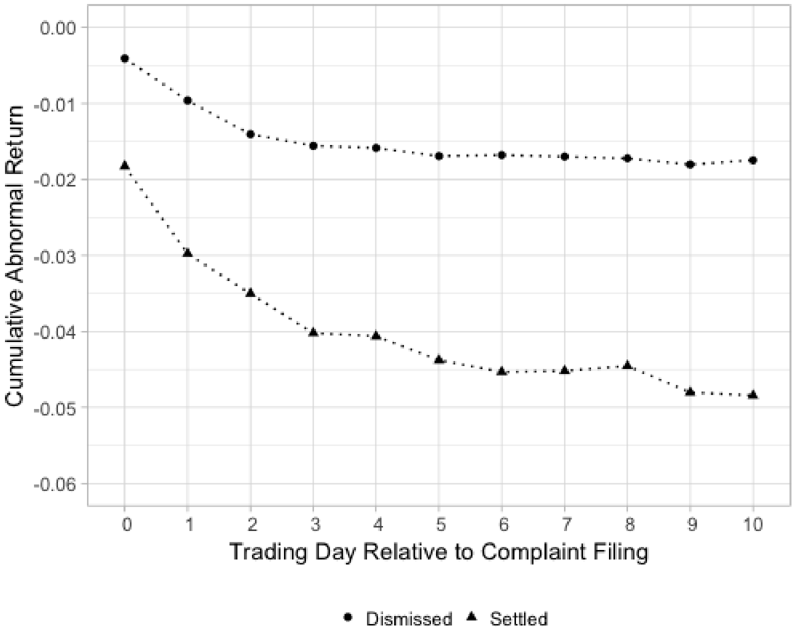

Figure 1: Cumulative Abnormal Return for First-Filed Cases Based on Eventual Outcome

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 estimated using a three-factor Fama-French model over days $t = -300$ to $t = -50$.

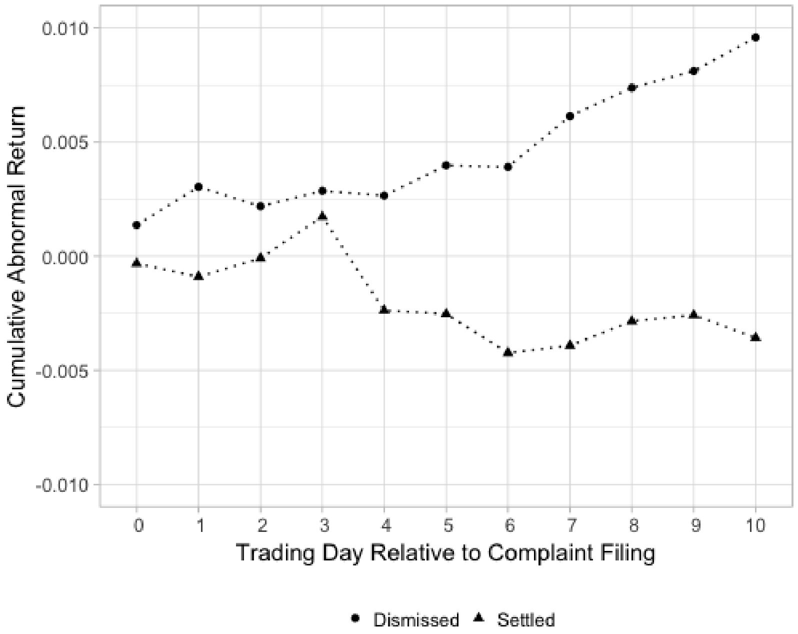

Figure 2: Cumulative Abnormal Return for Consolidated Cases Based on Eventual Outcome

Note: The figure shows the cumulative abnormal return over event days [0,+10] for consolidated cases based on whether those cases are actually settled or actually dismissed. The event is the filing of the complaint. The abnormal returns are estimated using a three-factor Fama-French model over days $t = -300$ to $t = -50$.

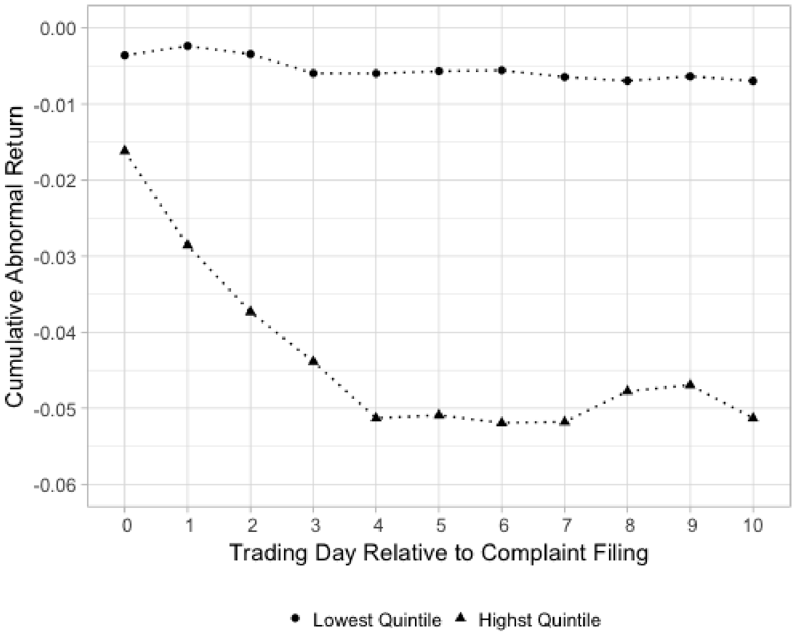

Figure 3: Cumulative Abnormal Return for First-Filed Cases in the Highest and Lowest Quintiles of Predicted Likelihood of Settlement

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 event is the filing of the complaint. The abnormal returns are estimated using a three-factor Fama-French model over days $t = -300$ to $t = -50$.

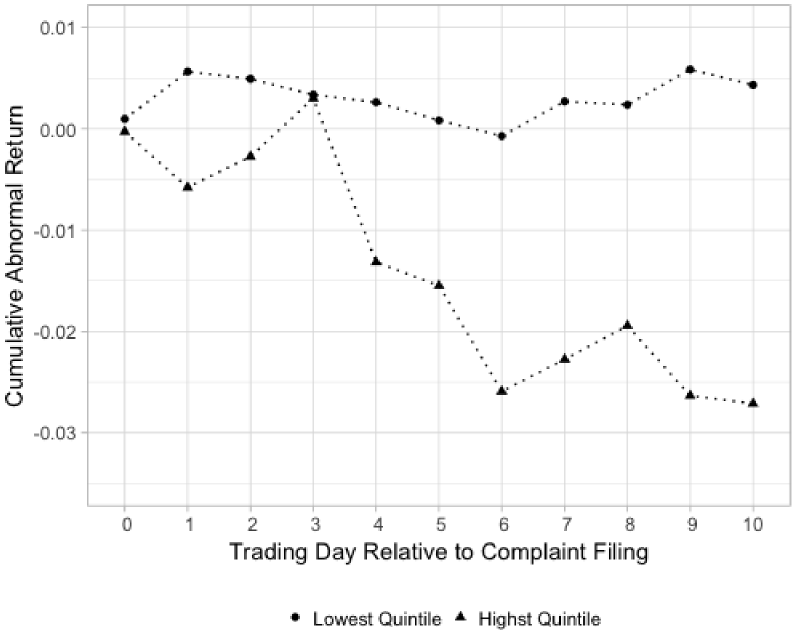

Figure 4: Cumulative Abnormal Return for Consolidated-Filed Cases in the Highest and Lowest Quintiles of Predicted Likelihood of Settlement

Note: The figure shows the cumulative abnormal return over event days [0,+10] for consolidated cases based on predicted likelihood of settlement. The predictions are derived from a random forest classifier that incorporates only text features. The event is the filing of the complaint. The abnormal returns are estimated using a three-factor Fama-French model over days $t = -300$ to $t = -50$.

Table 1: Summary Statistics for 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=2382, Baseline=56.7% Settled)

Classifier	Correct Settlements	Correct Dismissals	Incorrect Settlements	Incorrect Dismissals	Accuracy
Naive Bayes	1233	1002	245	906	66.0%
SVM	885	1262	593	646	63.4%
Random Forest	1000	1376	478	532	70.2%
AdaBoost	981	1368	497	540	69.4%
XGBoost	969	1367	509	541	69.0%

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

Classifier	Correct Settlements	Correct Dismissals	Incorrect Settlements	Incorrect Dismissals	Accuracy
Naive Bayes	809	670	541	809	62.1%
SVM	894	564	456	468	61.2%
Random Forest	1031	505	319	527	64.5%
AdaBoost	1047	515	303	517	65.6%
XGBoost	1009	576	341	456	66.5%

Panel C: All Cases Predictions (N=5768, Baseline=51.0% Dismissed)

Classifier	Correct Settlements	Correct Dismissals	Incorrect Settlements	Incorrect Dismissals	Accuracy
Naive Bayes	2179	1637	649	1303	66.2%
SVM	1816	1956	1012	984	65.4%
Random Forest	2136	1940	692	1000	70.7%
AdaBoost	2132	1947	696	993	70.7%
XGBoost	2028	1979	800	961	67.5%

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

Panel A: First-Filed Case Predictions (N=3270, Baseline=54.9% Dismissed)

Classifier	Correct Settlements	Correct Dismissals	Incorrect Settlements	Incorrect Dismissals	Accuracy
Random Forest	944	1262	532	532	67.5%
AdaBoost	917	1248	559	546	66.2%

Panel B: Consolidated Case Predictions (N=2380, Baseline=56.7% Settled)

Classifier	Correct Settlements	Correct Dismissals	Incorrect Settlements	Incorrect Dismissals	Accuracy
Random Forest	991	493	358	538	62.4%
AdaBoost	1004	490	345	541	62.8%

Appendix

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 A1 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 A1: 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