The Selection of Claims in Securities Arbitration for Settlement: A Text-Based Analysis

Stephanie Denis, Matthew L. Kozora, Minwen Li,
and Jonathan Sokobin[1]

December 31, 2021

ABSTRACT

This paper develops text-based measures of case strength and similarity using Statements of Claims filed in securities arbitration concerning investments in Puerto Rico municipal bonds to predict arbitration outcomes. We find that stronger customer claims and claims more similar to previous claims are more likely to settle than result in an award. These claims are also associated with higher total customer payout, including both settlements and awards. Our findings offer unique insights into the role of settlements as an alternative method of dispute resolution, and the selection bias associated with claims which result in an award.

Keywords: Securities Arbitration, Settlement, Statement of Claim, Textual Analysis

---

[1] Denis, fadenis@gmail.com, Kozora, matthew.kozora@finra.org, Li, minwen.li@finra.org, and Sokobin, jonathan.sokobin@finra.org, Office of Chief Economist of Financial Industry Regulatory Authority (FINRA). Ms. Denis' contributions to this paper occurred during her tenure in the Office of the Chief Economist. We thank Robert Garrison, Antony Gaetani, and An He for valuable technical support and research assistance; Paul Rothstein, Dror Kenett, and seminar participants at the FINRA Office of the Chief Economist brownbag seminar series and the FINRA Economic Advisory Meeting (2021) for their comments. The ideas and opinions expressed in this article are the authors' and should not be interpreted as reflecting the views of either FINRA or its staff.

1. Introduction

Securities arbitration is the most frequently used method to resolve disputes between an investor and their broker-dealer (FINRA, 2018). In securities arbitration, parties agree to abide by the decisions of a neutral third-party. Potentially more important than arbitral decisions (awards), however, is the role of the arbitration proceedings to faciliate settlements. Indeed, the economic literature has long argued that settlements negotiated by bargaining tend to be superior to awards (Crawford, 1979).[2] Consistant with this observation, the majority of disputes in securites arbitration are settled rather than decided by arbitrators.[3] Despite its importance, few studies investigate settlements in arbitration and related settings.

In this paper, we investigate the decision to settle a dispute in securities arbitration. We first describe FINRA's Dispute Resolution Services Forum (the Forum), which operates the largest securities arbitration forum in the U.S. and serves as the crucible for our empirical tests. We then present a simple model based on the work of Priest and Klein (1984) to motivate our empirical tests. In the model, two parties with imperfect information make rational decisions to maximize their expected utility of award. The model predicts that stronger customer claims, i.e., those claims where customers are better able to evidence the liability of industry parties, are more likely to settle than weaker customer claims. The model also predicts that parties are more likely to settle when

---

[2] There has been a concerted effort by economists to assess the welfare implications of arbitration. Crawford (1979) points out that settlements negotiated by bargaining tend to be superior to arbitral decisions because participants are likely to know more about their own and each other's preference than the arbitrator does. It is also more likely that they will accept a negotiated settlement and try to make it work.

[3] According to the statistics published by FINRA, approximately 70 percent of customer claims were resolved by a direct settlement by parties or settled via mediation in 2020. See https://www.finra.org/arbitration-mediation/dispute-resolution-statistics.

customers bring claims more similar to previous claims, and thus more information is available to the parties which describes a potential settlement or award.

Next, we empirically investigate these expected relationships. We apply natural language processing (NLP) methods to analyze the Statements of Claim (SOCs) filed by customers in the Forum. In line with the textual analysis literature (Hanley and Hoberg, 2010), we construct three empirical measures to describe the information content of an SOC: the total number of meaningful words (document length), its average similarity to previously filed SOCs (document similarity), and the percentage of negative words to total meaningful words (negative tone). We use document length as a proxy for the strength of the customer claim, and document similarity as a proxy for the amount of information available which describes a potential settlement or award in the model.[4] We also examine negative tone because previous literature (see Merkl-Davies and Brennan (2007) for an extensive review of the literature) suggests that document tone can be employed for the purpose of strategic disclosure and impression management, and thus can have an impact on the decision to settle.[5]

We restrict our sample to claims concerning investments in Puerto Rico municipal bonds filed and closed through the Forum from January 2014 to September 2020.[6] These cases represent 38 percent of all customer claims filed and closed in the Forum during this period. The sample

---

[4] Existing studies on arbitration outcomes largely ignore the effect of case strength and similarity. A few studies use indirect measures for case strength such as a request for expungements by respondent or a request of punitive damage by claimant (Choi, Fisch, Pritchard, 2010). These measures can be relatively noisy. For example, some lawyers will request punitive damages in every case, while others never do.

[5] See Section 4.2 for a more detailed discussion on the textual measures.

[6] Customer disputes relating to Puerto Rico municipal bonds were typically filed after 2014 when Puerto Rico municipal bonds were downgraded to non-investment grade due to weakening economic conditions. These complaints were filed over six years. This is an example of how customer claims relating to the same underlying events may be brought to the Forum over an extended period of time.

gives us a clean setting to test for the effects of claim strength and similarity while limiting the potential differences in the characteristics of the claims, and the securities and underlying events which led to the claims. Notably, all claims in our sample relate to investments in similar financial products and concentrate in few brokerage firms.[7] We still find, however, a large amount of variation among SOCs in our sample. The average similarity of SOCs in our sample is 11.4 percent. Although the SOCs drafted by the same attorney have similarities that are on average 28.5 percentage points higher than those drafted by a different attorney, we observe substantial variation in the SOCs drafted by the same attorney.[8,9]

Consistent with our model's predictions, we find longer SOCs are associated with a higher likelihood of settlement and a higher total customer payout, and more similar SOCs are associated with a higher likelihood of settlement.[10] We also find that SOCs with a more negative tone are associated with a lower likelihood of settlement and a lower total customer payout. These last results are consistent with the notion that a customer may employ a negative tone in an attempt to sway arbitrators or other parties to a case, and potentially in place of additional facts or evidence

---

[7] Five firms account for 86 percent of our sample cases and one firm alone accounts for 60 percent. In addition, our sample concentrates in allegations involving security recommendations (e.g., breach of fiduciary duty or suitability) than other issues (e.g., fees charged in error, excessive trading). Table 2 provides a detailed discussion of our sample by year, firm, and controversy types.

[8] We identified five cases in our sample where customers were represented by securities arbitration clinics affiliated with law schools. We did not identify any cases in our sample where customers were represented by non-attorney representatives (NARs).

[9] Such a signature effect may be subject to various interpretations. For example, it can indicate that arbitration attorneys tend to select cases that are similar in nature (i.e., a clientele effect). It may also capture attorney-specific knowledge or expertise (i.e., a certification effect). There is also a possibility that attorneys may have incentive to conform to the writing styles in previously arbitrated cases.

[10] To conduct this analysis, we take advantage of the obligation of firms to report, on behalf of their associated persons, the total amount of monetary and non-monetary compensation resulting from arbitrations including both settlements and awards. Section 4.3 provides a full discussion of the disclosure obligation and our methodology to construct the payout variables.

supporting the case.

The relationship between our textual measures and measures of arbitration outcome are economically significant. In our sample, the median total payout is $100,000, and the median total payout-to-claim ratio is 28 percent. A one standard deviation increase in document length (document similarity) is associated with a $25,740 ($19,860) increase in customer payout, a magnitude of 25.7 percent (19.8 percent) of the sample median. The increase also leads to a 7.9 percent (21.8 percent) increase relative to the median total payout-to-claim ratio.

Our evidence is robust to a comprehensive battery of tests. First, we consider alternative textual measures and use of the last SOC filed instead of the longest SOC filed. Second, we consider alternative measures for total payout and total payout-to-claim ratio. We also consider alternative specifications to the model. Finally, our results are robust to an alternative sample involving only those cases with assertions of fiduciary duty or suitability violations.

Our study contributes to the academic literature in several significant ways. First, our study adds to the law and economic literature that investigates the selection of disputes for litigation. The majority of existing theories predict a selection effect for litigation (e.g., Priest and Klein, 1984; Hylton, 1993; Shavell, 1996; Lee and Klerman, 2016). They show that there are material differences between litigated cases and cases that settle under a variety of conditions. Empirical work to test these theories, however, yields inconclusive results.[11] In this paper, we adapt the Priest

[11] The empirical work typically focuses on the plaintiff trial win rate (e.g., Eisenberg, 1990; Kessler, Meites, and Millel, 1996). Only a few studies include information on settled cases that allows for a direct comparison between settled and litigated cases. For instance, Klerman (2012) examines privately-prosecuted criminal cases in the thirteenth century when judges took jury verdicts even in settled cases, and Studdert and Mello (2007) examine insurance files in medical malpractice cases. Both papers find that strong cases are more likely to settle than result in litigation. Helland, Klerman, and Lee (2018), however, find no significant selection effects after comparing the distribution of settlement versus adjudicated damages using New York "closing statement" data.

and Klein (1984) model to securities arbitration, and find consistent empirical results showing that the cases that settle differ from those that result in an award in predictable ways.

We also contribute to the academic literature by shedding light on the potentially large role of settlement in arbitration. Our analysis is aided by access to the SOCs made by customers in the Forum. Among other things, the SOCs permit us to identify the damages claimed by customers. This information permits us to measure the ratio of payout to claims for the majority of cases in our sample. Another internal source permits us to accurately identify the manner in which a case closes, if not by award. These data sources permit us to accurately identify and measure arbitration outcomes, and are not available to other researchers.

While existing economic theories indicate that voluntary settlement is superior to arbitral decisions,[12] almost all of the empirical evidence available relates to arbitration cases that resulted in an award (e.g, Bloom and Cavanagh, 1986; Bloom, 1986). By presenting systematic evidence concerning settlement, our study helps contribute to the understanding of its role. Our study also demonstrates that a sample limited to adjudicated outcomes in arbitration may be subject to a significant selection bias.

For securities arbitration in particular, researchers have used awards to investigate the fairness of arbitration (Kondo, 2007; Egan, Matvos and Seru, 2020), arbitrator background (e.g, Choi, Fisch, and Pritchard, 2010, 2014), punitive damages (Choi and Eisenberg, 2010), arbitrator

---

[12] The majority of the theory literature applies economic bargaining models in the context of labor disputes. See, Kuhn (2009), for a review of the literature. Related laboratory experiments center on whether an arbitration method encourages settlement and what initially causes bargainer disagreement. Evidence suggests that bargainers may resort to arbitration because they are overly optimistic about the award that he or she will receive from the arbitrator (e.g., Faber and Bazerman, 1989; Dickson, 2005), or because the parties hold unequal information in arbitration (e.g., Farmer and Pecorino, 1998; Pecorino and Boening, 2001). Further, Farber, Neale, and Bazerman (1990) document risk aversion of participants in experiments, and find that negotiated settlements seem to favor the less risk-averse bargainer.

recommendations to expunge customer dispute information from an individual's record (Honigsberg and Jacob, 2021), and the effect of securities laws on broker-dealer liability (Kozora, 2017). Our analysis suggests that although potentially informative, these studies may be limited by the representativeness of the samples and the ability of researchers to adequately control for the selection of claims for securities arbitration.

Finally, we add to the thriving academic literature using textual analysis to examine the informativeness of written disclosures to test theories in economic, finance, and law. Hanley and Hoberg (2010, 2012) find that strong disclosure in IPO prospectuses reduces IPO underpricing and hedges against future lawsuits. Fresard, Hoberg, and Phillips (2020) and Hoberg and Phillips (2010) use textual analysis to test theories of mergers and acqusitions. Loughran and McDonald (2014) show that firms using plain English have better governance.[13] In our application, we find robust evidence that information content of customer claims in securities arbitration can predict the likelihood of settlement and customer payout. While our empirical analysis is focused upon a single class of cases relating to the failure of Puerto Rican municipal securities, the variation in the claims, location of the claimant, and other characteristics associated with the arbitration leads us to believe that our findings are informative for securities arbitration in general.

The rest of the paper is organized as follows: Section 2 provides background information describing the Forum; Section 3 presents our simple model to motivate the empirical analysis; Section 4 presents the data, variable construction, and descriptive statistics; Section 5 presents

[13] In other context, papers such as Hanley and Hoberg (2019) and Baker, Bloom and Davis (2016) use word content from disclosure and newspaper coverage to capture emerging risk in the financial sector and economic policy uncertainty. You and Zhang (2009) and others examine information embedded in 10-K filings and investors' reaction to such information. Correa et al (2020) analyze the relation between the sentiment conveyed by financial stability reports published by central banks and the financial cycle.

empirical evidence describing the sources of SOC content; Section 6 presents empirical evidence relating the text-based measures and arbitration outcomes; Section 7 describes robustness tests; and Section 8 concludes.

2. FINRA Arbitration

FINRA is a self regulatory orgnaization authorized by the U.S. Congress and under supervision by the U.S. Securities and Exchange Commission. It is responsible for the broker-dealer industry with the mission to protect investors and ensure market integrity. As part of its function, FINRA operates the largest securities dispute resolution forum in the U.S. Broker-dealers often require customers to enter into agreements to arbitrate disputes arising from the services provided to such customers. Under FINRA rules, arbitration in the Forum is required if there is a written agreement requiring FINRA arbitration or if it is requested by the customer.[14] FINRA rules, however, do not require such agreements, nor do they preclude customers from pursuing relief in state or federal courts. Annually, customers file over two-thousand new cases against firms and associated persons (i.e., industry parties) in the Forum.[15]

A customer case originates when a customer experiences a loss and believes that he or she has been treated improperly. The customer can register a complaint with the firm. The parties to

---

[14] *See* FINRA Rule 12200; *see also* Regulatory Notice 16-25 (July 2016) (reminding member firms that customers have a right to request arbitration at FINRA at any time and do not forfeit that right under FINRA rules by signing any agreement specifying another dispute resolution process or venue). Even with a predispute arbitration agreement, member firms and customers may elect, by mutual consent, to resolve their disputes in a forum other than at FINRA, such as at a private arbitration forum or by civil litigation, *after* a dispute has arisen between the parties. Similarly, if a written agreement to arbitrate at FINRA does not exist and the customer does not request FINRA arbitration, the parties to a dispute may proceed to agree to resolve their disputes at a private arbitration forum or in civil litigation.

[15] For annual statistics, see Footnote 3.

the complaint may then reach a resolution, which can involve compensation (i.e., settlement) to the customer. If the parties do not directly settle the dispute, then the customer can initiate an arbitration by filing an SOC with the Forum. The Forum then serves the SOC on the firm and/or associated person identified in the SOC, and these parties may respond by filing a Statement of Answer (SOA). In customer cases, the customers (i.e., plaintiffs) are known as claimants, and the industry parties (i.e., defendants) are known as respondents. A customer case can relate to one or more disputes, and customers can name one or more firms and associated persons in the complaint.

The arbitration proceeds as follows. First, the two parties jointly select the arbitrator or panel of arbitrators to oversee the proceedings. At this time, parties also engage in discovery where they exchange documents and other information, and attend any prehearing conference where the arbitrator sets process deadlines, addresses other preliminary matters, and facilitates any ongoing discovery-related issues. Next, the parties attend evidentiary hearings where they can provide oral testimony, submit documentary evidence, and cross-examine opposing witnesses. Finally, if the parties still have not resolved the dispute by means of a direct settlement, then the arbitrator or panel considers all evidence and decides the case.[16] A verdict or judgement in arbitration is known as an award. Figure 1 presents a timeline for a customer dispute.

See Figure 1

[16] A customer claim in the FINRA Forum can close by means other than by settlement or award. For example, approximately three percent of customer claims in our sample are withdrawn. A customer claim in the Forum can also result in both a settlement and award. This may occur if a claim includes more than one complaint or is brought against more than one industry party. A customer case involving multiple customers or industry parties may result in more than one outcome. For example, a customer case may result in a partial settlement and partial award if one or more parties settle, and the arbitrator or panel of arbitrators decides the remaining disputes.

3. Theoretical Framework

In this section, we develop a simple model to motivate the empirical tests that follow. The model describes the process for a customer to arbitrate a claim against an industry party, and makes predictions regarding when parties may settle or arbitrate the dispute. The model incorporates elements from the Forum but may be generalizable to other dispute resolution forums depending on the policies and procedures of those forums.

The intuition of the model is that parties have imperfect information to assess a potential award in arbitration. In expectation, neither party has an informational advantage to estimate a potential award. There is learning by both parties as the arbitration unfolds and more information becomes available. As the information between the two parties becomes more common, the parties' estimates of award will converge and the likelihood of settlement will increase. Industry parties also stand to lose more than just the award, with additional stakes which increase with the size of the award. These additional stakes increase the likelihod parties settle the dispute when the customer brings a stronger claim to the Forum.

Priest and Klein (1984). In their base model, the authors argue that parties will settle cases where the outcome of the litigation is more certain (i.e., the legal standard favors either the plaintiffs or defendants) but will litigate cases where the outcome of the litigation is less certain (i.e., the legal standard favors neither party), and the success rates of either party will tend toward fifty percent. In an extension of their model, the authors also argue that there may be deviations from fifty percent in favor of the parties who may incur stakes in addition to the award.[17]

3.1 Information Flow

[17] Lee and Klerman (2016) show that this second prediction holds under a broad range of assumptions.

Parties have an initial endowment of information prior to and at the filing of an arbitration relating to the activities that underlie the claim. As the arbitration progresses, parties update their information as new information becomes available, such as when the parties engage in discovery or attend evidentiary hearings. We assume that although the information available to parties may differ, neither the customer nor the industry party has information which more closely reflects the potential award.[18] This assumption implies that neither party will have an advantage with respect to the information available to them when deciding whether to settle or arbitrate a claim.

Information describing settlements and awards from previous claims may also be available to both parties. When the parties have not yet engaged in discovery, previous settlements and awards may suggest the information that may become available, and in particular when the cases relating to the previous cases are more similar to the current case. Parties can use this information to estimate a potential award.[19] Arbitrators, however, will determine an award based solely on the information brought to the Forum as part of the arbitration.

3.2 Award Estimates

---

[18] Other studies model the decision to settle a dispute in litigation as a negotiation where one party has an advantage to assess a potential litigation outcome. Examples of these studies include Bebchuk (1984), and Reinganum and Wilde (1986). In general, for the claims in our empirical sample, we do not believe that one party would necessarily have an advantage to assess a potential award. Factors that could influence an award such as the underlying events relating to the sale of securities and the identified risk -preferences of the customer should be known by both parties.

[19] The assumption that both parties have access to information describing previous settlements and awards is consistent with the potential information available for the cases in our empirical sample. The summaries of cases resulting in an award are publicly available and provide a description of the claim and its outcome, and settlements greater than de minimus thresholds are generally available. See Section 4.3 for a discussion of the availability of settlement information.

We assume that customers and industry parties are risk-neutral,[20] and form rational expectations of award. Parties estimate awards at each stage of the arbitration process, and base these estimates on the information available at the time. Parties also interpret the same information with a similar underlying independent random process with mean zero error. Although it is likely that the industry party has additional experience in the Forum, and may therefore have greater ability to interpret information as it relates to a potential award, customers may retain representation who specialize in securities litigation or otherwise have experience in the Forum.[21] Arbitrators will interpret information similar to customers and industry parties.

Although on average the parties will interpret the same information in a similar fashion, their estimates of potential awards may differ. One reason may be that parties may have different information prior to engaging in discovery. But, even with the same information, differences in interpretation by either party may cause differences in these estimates. We denote the customer's estimate of award with A_C, and the industry party's estimate of award with A_I.

Industry parties may incur more loss than just that relating to the immediate arbitration. These additional stakes, for example, may relate to a potential loss in reputation. The additional stakes may also result in an increase in future claims after other customers observe an award.[22] We

---

[20] We could instead assume that parties are uniformly risk-averse, or exhibit risk preferences consistent with prospect theory or other behavioral models (Korobkin and Guthrie, 1994; Rachlinkski, 1996; and Korobkin and Ulen; 2000). Predictions of arbitration outcomes may differ under these alternative assumptions. We believe, however, that in general our hypotheses would still hold.

[21] This assumption is consistent with the overwhelming majority of customers in our empirical sample who retained legal counsel.

[22] We note that economically the additional stakes of industry parties are equivalent to a cost. We separately identify the additional stakes in the model for consistency with Priest and Klein (1984).

assume that these additional stakes, k, are an increasing function of A_I ($k(A_I)$). This implies that in expectation, industry parties incur greater additional stakes as the size of the award increases.

Priest and Klein (1984) also incorporate estimates of award in their model, and represent an award estimate as the product of (1) the assessed probability of defendant (or respondent) liability verdict based on the legal standard and (2) the expected judgement should the defendant (or respondent) be determined liable. In this model, we do not similarly bifurcate award estimates between the expected likelihood of a verdict in favor of one party and the expected magnitude of the judgement conditioned on the verdict. FINRA arbitration is an equitable forum, and arbitrators are not strictly bound by legal precedent or statutory law. Although the Forum guides arbitrators to first determine whether an industry party is liable and then the appropriate remedy, information describing the underlying process which led to the award and the basis for that amount are not typically available. In most cases, therefore, we are only able to observe the total award, which embeds the ex post measure of liability, and not the two separate components as in the Priest and Klein (1984) model.

3.3 The Settlement Decision

We follow Priest and Klein (1984) and incorporate the condition for settlement as established by Landes (1971), Posner (1973), and Gould (1973). Customers and industry parties will base their decision to settle by comparing the financial value from settling a claim to the anticipated value from arbitrating the claim.[23] Let S_C represent the minimum settlement demand of the customer and S_I represent the maximum settlement offer of the industry party. These values

[23] Customers and industry parties may also base their decision to settle a dispute in arbitration on other factors not relating to remuneration (e.g., to testify in a hearing against the opposing party). Legal counsel, however, may decrease the influence of behavior on arbitration outcomes (Korobkin and Guthrie, 1994).

reflect the net anticipated gain or loss from arbitrating the dispute after compensation for the costs incurred to settle the dispute. S_C and S_I are represented as

$$S_{I,t} = (A_{I,t} + k(A_{I,t}) + C_{A,I,t}) - C_{S,I,t} \quad \text{and} \quad S_{C,t} = (A_{C,t} - C_{A,C,t}) + C_{S,C,t} \tag{1}$$

where at time t $A_{C,t}$ ($A_{I,t}$+ $k(A_{I,t})$) represents the present value of the customer's (industry party's) estimate of the gain (loss) from the arbitration award, $C_{A,C,t}$ ($C_{A,I,t}$) represents the present value of the sum of current and future costs to arbitrate the claim, and $C_{S,C,t}$ ($C_{S,I,t}$) represents the current costs to settle the claim.

The parties may settle a claim if there is a range of settlement amounts which leaves both parties better off. This range exists if the maximum amount for which an industry party may offer is greater than the minimum amount for which a customer may accept ($S_{I,t} > S_{C,t}$). In these instances, the difference between the two settlement amounts reflects the range of settlement amounts which satisfy the condition, and can be represented with the following inequality.

$$(A_{I,t} + k(A_{I,t})) - (A_{C,t}) > (C_{S,I,t} + C_{S,C,t}) - (C_{A,I,t} + C_{A,C,t}) \tag{2}$$

or

$$(A_{I,t} + k(A_{I,t})) - (A_{C,t}) > C_{S,t} - C_{A,t} \tag{3}$$

where $C_{S,t} = C_{S,I,t} + C_{S,C,t}$ and $C_{A,t} = C_{A,I,t} + C_{A,C,t}$.

As of the filing of the claim, when there has been no formal exchange of information between parties, the inequality suggests that customers will file claims in arbitration when they

anticipate a relatively larger award than the industry party.[24] Following the filing of the claim, the minimum amount for which a customer is willing to settle will increase with their estimate of the anticipated award. To the extent such an estimate is dependent on the customer's ability to demonstrate the liability of the industry party, the minimum amount for which a customer is willing to settle positively relates to the strength of his or her claim.

Hypothesis 1: *For a given level of customer loss, stronger customer claims result in a higher payout than weaker customer claims.*

All else equal, a settlement is also more likely when the industry party anticipates a strong customer claim. In these instances, because of the asymmetric stakes, the likelihood that $A_{I,t}$+ $k(A_{I,t})$ is large relative to $A_{C,t}$ is higher and a settlement is possble which leaves both parties better off.[25]

Hypothesis 2: Stronger customer claims are more likely to settle than weaker customer claims which are more likely to result in an award.

---

[24] Parties may directly settle the dispute prior to the filing of a claim or after each stage of the arbitration. We focus this discussion, however, on the time period following the filing of the claim to frame the empirical analysis. These same predictions, however, would also apply to the time period before the filing of the claim.

[25] Without the asymmetric stakes of the industry party, the otherwise symmetric nature of the model would suggest that the strength of the customer case would not factor into whether a case settles or goes to award, and arbitration awards would favor neither the customer nor the industry party. Other models which instead incorporate asymmetric information between parties (e.g., Bebchuk, 1984; Hylton, 1993) also predict stronger customer claims are more likely to result in settlement.

The inequality also suggests the potential importance of the availability of information to the likelihood of settlement. Regardless of the strength of the customer claim, the more similar the claim to previous claims, and thus the more information available to both parties describing a potential settlement or award, will result in smaller differences in award estimates. The smaller differences in award estimates decrease the likelihood that the customer's award estimate remains large relative to the industry party's estimate and the inequality fails to hold.

Hypothesis 3: Claims more similar to previous claims are more likley to settle than claims less similar to previous claims.

4. Data and Descriptive Statistics

4.1 Sample and Data Sources

Our sample consists of 3,235 customer claims filed and closed in the Forum between January 2014 and September 2020 pertaining to investments in Puerto Rico municipal bonds. These cases account for 38 percent of all customer cases filed and closed during the sample period. We identify these cases using information from the SOCs customers file in the Forum. We also use the information from SOCs to describe the size of the claim, the rule violations customers assert, and the legal representation of a party. From these documents, we construct textual variables on document length, negative tone, and similarity. We also use information from other internal sources to describe the manner in which a case closes, the arbitrators appointed to the panel, and the hearing location. To describe the total payout relating to a case, we combine information from the arbitrations which close by award with information from the Central Registration Depository

(CRD).[26]

4.2. Textual Variables

Existing research investigating arbitration outcomes is limited by an inability to directly control for case strength or similarity. We are able to directly control for these characteristics by applying NLP methodology to analyze the information content of SOCs. This section describes our text-based measures.

4.2.1. Textual Analysis

We start by obtaining Forum documents and identifying the SOCs.[27] When multiple SOCs are filed for the same case, we analyze the longest one. We also analyze the most recent SOCs for robustness. Prior to applying the textual analysis, we apply several filters to the content of the SOCs.

- We delete numbers, tables, graphs, and non-meaningful words such as prepositions, articles, conjunctions, and pronouns (Loughran and McDonald, 2011).
- We exclude exhibits from the main texts as these exhibits often relate to different subjects, for example, email correspondence, contracts, prospectuses, and regulatory policies or proceedings.[28]
- Finally, we remove sparse words that appear less than 100 times across all the

---

[26] Information relating to cases that resulted in an award can be obtained through FINRA Arbitration Awards Online (https://www.finra.org/arbitration-mediation/arbitration-awards). Information describing arbitration awards include a summary of the claim and the award decision. Information from CRD is publicly available through FINRA BrokerCheck (https://brokercheck.finra.org/).

[27] Other types of forum documents submitted by the customers may include submission agreement, motion to compel, motion to extend, etc. A typical SOC in our sample ranges from several pages to a hundred pages. It describes the jurisdiction, parties, facts and allegations of potential misconduct and alleged damages.

[28] Our main findings are robust to the alternative sample that only includes SOCs without exhibits.

documents. This step helps mitigate the effect of potential errors generated during the Optical Character Recognition (OCR) process.[29,30]

See Figure 2

Figure 2 displays the cluster of the top 100 meaningful words used in the final sample. Words that are bigger and bolder appear more frequently in the sample. According to the figure, "UBS," "Puerto," "Rico," "funds," "bonds," and "claimant" are among the most frequently used words, which is consistent with the characteristics of the cases in the sample.

We develop three types of textual variables. Our first textual variable, *Length*, is defined as the total number of meaningful words in an SOC. We use this measure to proxy for case strength in our model.[31] We conjecture that a customer who submits a longer SOC is better able to evidence the liability of industry parties and is therefore likely to have a stronger case. This is based on the belief that customers are inclined to disclose information material to the case in the SOC as it is the first description of the case that an arbitrator or panel may review.

Several factors may cause finding the expected relationship between *Length* and case outcomes to be more difficult. First, customers may initially withhold information as a litigation strategy. Second, as found by Hancock et al. (2007) in the context of communication via text messages, length may instead be associated with deception (although opposite evidence has been

---

[29] Our documents may be prone to OCR errors because the majority are manually scanned files.

[30] We conduct several robustness checks using different cutoffs for sparse words (e.g., no cutoff or a cutoff that requires a word to appear 30 times or more in total) and obtain similar results. After implementing the above procedures to remove non-meaningful content, our final sample of documents has a vocabulary of 5,800 unique meaningful words.

[31] One may argue that the length of an SOC can also proxy for case complexity. In our sample, however, all customer claims relate to investments in Puerto Rico municipal bonds. Therefore, to a large extent, similarity in the underlying financial product and activities that have led to the claim helps alleviate such a concern.

found in other contexts, e.g., Toma and Hancock, 2012). Finally, our measure may not be able to capture language subtleties such as whether the words represent a fact or an opinion, or whether the fact itself is relevant.[32]

Our second textual variable, *Negative Tone*, is computed as the percentage of negative words to all meaningful words. We use the negative word dictionary developed by Loughran and McDonald (2011). Their negative word dictionary reflects tone in financial contexts. For instance, appearances of these words in company Form 10-Ks are associated with negative future stock returns, fraud, and various other bad financial outcomes. This dictionary is particularly useful in our context because customers' arguments often center on investment activities and financial outcomes.[33] Examples of negative words used in our cases include "damages," "fail," "violation," and "imprudent."

As mentioned earlier, existing studies show that managers adopt a specific tone in corporate disclosures (e.g., earnings releases, annual reports, earnings conferences calls, and CEO letters) to strategically influence analysts, investors, media, and other stakeholders' expectations. Consistent with these findings (e.g., Huang, Teoh, Zhang, 2014) and the relevant theories on impression management (Schlenker,1980; Merkl-Davies and Brennan, 2007), we conjecture that customers engage in impression management by adding additional color, potentially in lieu of factual information, in an attempt to sway arbitrators and other parties. If true, then we should find a negative relationship between our measure of negative tone and customer payout. Alternatively, a

---

[32] In a similar vein, it is possible that our bag-of-word approach may not be able to capture semantic meaning of words used in customers' statements.

[33] We also construct an alternative negative tone measure using the Harvard Psychosociological Dictionary (Loughran and McDonald, 2010), even though this dictionary is not designed to capture negative tone specifically in financial context. Our results indicate that the alternative negative tone measure leads to similar, if not stronger, results.

negative tone may simply reflect a personal trait or style of a customer (e.g., pessimism), and therefore should have no direct correlation with customer payout (Amicis, Falconieri, and Tastan, Forthcoming). Finally, there is a possibility that customers embed a negative tone when disclosing the facts and evidence indicating industry party liability. If so, then we should find a positive relationship between our measure of negative tone and customer payout.

Finally, following the previous literature (e.g., Hanley and Hoberg, 2010; Hoberg and Phillips, 2016; Loughran and McDonald, 2020), we develop two similarity measures: *Pairwise Similarity* and *Avg. Similarity*. We characterize each SOC by a word vector, which consists of its loadings (i.e., frequency counts) on each unique word in the full sample. *Pairwise Similarity* is then defined as the cosine similarity between any two SOC word vectors. This is a numerical variable ranging from zero to one, with a value of zero indicating that the documents are entirely different with no words in common. Appendix A includes a detailed description of the main variables used in our analysis. We compare a specific document to all other previously filed documents and calculate the average similarity score by averaging across its pairwise similarities with these other documents. We call the estimates the case-level average similarity (i.e., *Avg. Similarity)*.

An SOC with high average similarity suggests the current claim is similar in nature to previous claims, and therefore more information from previous cases may be available to estimate a potential award. As the model suggests (Hypothesis 3), more information which is common among parties to estimate a potential award should increase the likelihood to settle.[34]

[34] Alternatively, an SOC with high average similarity may reflect uninformative, "boiler-plate" statements, or an attorney's attempt to copy the SOC content of previous, dissimilar cases. If true, then it may not relate to the amount of common information among parties or a higher likelihood to settle.

4.2.2. Characteristics of the Text-Based Variables

Panel A of Table 1 presents the distribution of our text-based variables. An average SOC has 4,215 meaningful words. Around 8 percent (or 330) of the words are negative. The 3,235 documents form approximately 5.2 million unique document pairs. The average similarity of these pairs is 11.4 percent with a standard deviation of 15.2 percent. The average of the case-level average similarities is 8 percent with a standard deviation of 5.5 percent. Given that we limit our sample to a set of cases that are all related to Puerto Rican municipal bonds, these estimates suggest there is sufficient variation across case documents to test our hypotheses. In unreported analysis, we also examine whether there are potential changes in distribution across years. The results indicate no significant time trends.[35]

See Table 1

Panel B of Table 1 compares the averages of text-based measures by claim size and legal representation. We find that larger claims tend to have longer SOCs than smaller claims. Relative to when customers retain legal representation, SOCs when customers self-represent are shorter and use fewer negative words. SOCs filed by self-represented customers are also less similar to previously filed SOCs.

4.3. Customer Payout Variables

We measure the total payout from a customer claim in arbitration as the sum of all settlements and awards. A customer claim can result in both a settlement and award if it is brought to the Forum by more than one customer, often related, or brought against more than one industry

[35] In unreported analyses, we also examine the distribution of textual variables at the attorney level. In general, we find substantial variation in texts across SOCs prepared by the same claimant attorney. This suggests that our textual measures are not entirely driven by an attorney's specific writing style.

party. In these instances, parties may settle a portion of the claim and arbitrate the other portion which remains. We obtain settlement and award information from CRD, and award information from the Forum.[36] An award may include the loss attributable to the industry party (i..e, compensatory damages) as well as other damages such as punitive damages and attorney's fees. To obtain a single estimate from the two sources of information, we sum the settlements from CRD with the largest estimate of award between CRD and the Forum.[37]

The settlement and award information from CRD derives from firm disclosures on the individual uniform forms – Form U4 and Form U5. FINRA rules require firms to submit these forms, on behalf of their associated persons, to either register or update the registration information of associated persons (Form U4) or terminate one or more of the registrations of associated persons (Form U5). The forms request information regarding customer complaints against the individual, including the resolution of the complaint such as through settlement, arbitration, or civil litigation.

Two aspects of CRD disclosures complicate our estimate of the total payout. First, not all settlements relating to an arbitration are available from CRD. By construction, firms do not disclose settlements on an individual uniform form if no individual is named as a respondent or the subject of the arbitration. In addition, firms are not required to disclose settlements on an individual uniform form if the amount is less than the de minimis threshold which is $15,000

---

[36] See Footnote 26.

[37] Arbitration awards from the Forum are publicly available. Firms reporting arbitration awards on an individual uniform form should therefore have information describing the award amount. There may be differences between the awards disclosed on an individual form and the awards available from the Forum and published online. These differences can relate to the damages included as part of an award. For example, payments associated with post-award interest may be reported on the individual uniform forms but would be outside the scope of the information that would be available to the Forum. To account for these potential differences, we estimate the award for an arbitration as the largest value disclosed among the individual uniform forms and from the Forum. Our results are robust to only using award information from the Forum.

during the sample period. Information describing settlements would also not be available if the record is expunged (Honigsberg and Jacob, 2021).

Second, multiple firms may disclose different settlements and awards relating to the same arbitration. Although the disclosure questions on the individual uniform forms refer to the total monetary amount, firms may not be privy, either directly or indirectly, to the total settlement if there is not one but multiple settlements. We assume that each unique settlement amount reflects a different part of the claim (i.e., a separate complaint), and estimate the total settlement as the sum of these unique values.

In general, these two factors do not affect a large percentage of the sample. For example, we can match 3,003 of 3,235 sample cases (92.8 percent) to at least one CRD disclosure. We should therefore be able to identify settlement amounts (greater than the de minimis thresholds) for most cases. Among the 3,003 cases that remain in the sample, no settlements were disclosed for 204 cases (6.8 percent). These cases may have resulted in a full or partial settlement less than $15,000. Finally, among these cases, only 76 cases (2.5 percent) involve more than one unique reported settlement amount.

We measure the total payout relative to the amount of damages claimed.[38] Where there is common agreement among parties in the dispute as to customer losses, the total payout ratio can be viewed as a proxy for the expected allocation of industry party responsibility for customer loss. The amount of damages claimed, however, may not reflect common agreement of parties as to customer losses. Parties may instead have different assessments as to the nature or size of the loss

[38] The claim amount is available for 2,957 of the 3,235 cases in the initial sample (92.1 percent). Customers may not specify a claim amount, and instead utilize discovery to determine the amount of damages. We replace missing claim amounts using median estimates from the other sample claims. In unreported tests, our empirical findings are robust to excluding these observations.

incurred by the claimant. In these instances, the total payout ratio would proxy for the expected allocation of industry party reponsibility for cuostmer loss with noise, and weaken our ability to find a relationship between the total payout ratio and our textual measures describing customers' claims. To the extent possible, we empirically control for other factors which may influence the total payout to customers that does not also relate to the strength of the customer claim (e.g., the identity of the arbitrators appointed and the hearing location).

4.4. Summary Statistics

Table 2 describes our sample of 3,235 cases. The results in Panel A of Table 2 show that the majority of the cases were filed between 2014 and 2018 with the number of cases evenly distributed across the five years. Our results in Panel B further indicate that the five firms most often named as a respondent were named in a large proportion of cases (2,799 cases or 86 percent). Among them, one firm alone was named as a respondent in 1,955 cases (60.43 percent).

Existing literature (e.g., Choi, Fisch, and Pritchard, 2010, 2014; Egan, Matvos, and Seru, 2020) has shown that hearing location can explain a significant portion of variations in customer awards. We report our cases by hearing location in Panel C of Table 2. The results indicate that 69 percent of the cases have hearing locations in San Juan, Puerto Rico. Other common hearing locations in the sample include New York City (NY), Boca Raton (FL), Atlanta (GA), and Miami (FL). These five locations relate to approximately 85.6 percent of the sample cases.

A single case typically involves multiple assertions of securities violations associated with the potential misconduct (i.e., *Controversy Type*).[39] Conceptually, the assessment of case strength

[39] Major types of violations in securities arbitration include breach of fiduciary duty, negligence, misrepresentation, failure to supervise, omission of facts, breach of contract, suitability, fraud, violation of Blue Sky Laws, manipulation, unauthorized trading, error-charges, elder abuse, margin calls, and churning.

and arbitration outcomes can vary depending on the specific type of underlying activity which led to the claim. Previous empirical work confirms that controversy types significantly predict customer awards in securities arbitration (e.g., Choi, Fisch, and Pritchard, 2010, 2014). We report controversies in Panel D of Table 2, and control for their effects in subsequent analyses. The most common allegations are breach of fiduciary duty or negligence (97.3 percent), failure to supervise (84.0 percent), misrepresentation or omission of facts (83.7 percent), breach of contract (82.6 percent), and suitability (79.7 percent).[40] Our evidence is largely consistent with prior work (e.g., Kozora, 2017), except that the Puerto Rico municipal bond cases in our sample tend to concentrate more in allegations involving security recommendations than other issues (e.g., fees charged in error, excessive trading).[41]

See Table 2

Table 3 reports summary statistics for the non-textual variables used in our analysis. Panels A, B, and C present the distribution of the variables on case characteristics, arbitration outcomes, and claimant legal representation, respectively. The median alleged damages (i.e., *Claim Size*) was $400,000. The overwhelming majority of customers were represented by attorneys, whereas very few (1.5 percent) represented themselves. There were 412 unique attorneys that represented customers in our cases. Among them, 143 attorneys represented more than one case. For these attorneys, the median attorney represented five cases. The pool of attorneys representing customers in more than one case permits us to examine the similarities of SOCs prepared by the same individual.

---

[40] See Appendix A for a description of various controversy types.

[41] For robustness, we exclude cases in which customers' allegations do not involve violations of fiduciary duty or suitability. The robustness results are reported in Appendix Table B3.

See Table 3

Approximately 91.5 percent of the cases were closed by settlement, whereas only 4.0 percent of cases resulted in an award after hearing (3.4 percent) or after review of the documents (0.6 percent).[42] Around 4.5 percent of cases were either withdrawn (3.5 percent) or closed by other means (1.0 percent). These numbers highlight the importance of examining settled cases to assess the efficiency of securities arbitration. At the median, the total payout was $100,000. The median total payout to claim ratio was 28.5 percent, indicating that a typical customer received approximately 29 cents per dollar claimed.

5. Sources of SOC Content

In this section, we explore the variations in SOC content and investigate potential factors that may have contributed to similar content across SOCs. This analysis can also help us better comprehend the similarity measure. We adopt an empirical methodology similar to Hanley and Hoberg (2010).[43]

See Table 4

We compare the similarity between any two SOCs by estimating ordinary least squares regressions. The dependent variable is *Pairwise Similarity*. As mentioned earlier, there are 5,230,995 unique document pairs among the 3,235 cases, and each pair reflects a regression observation. Table 4 presents the regression results.

---

[42] Compared with other customer disputes in the Forum, cases related to Puerto Rico municipal bonds tend to have a higher likelihood of settlement. Customers are also more likely to have legal representation in these cases.

[43] Hanley and Hoberg (2010) use textual analysis to assess the source of content in IPO prospectuses. The study develops methodologies to explore whether underwriters tend to use "standard" content across prospectuses drafted by them.

The first two explanatory variables identify whether the two cases were represented by the same claimant attorney (*Same Claimant Attorney*), or the same claimant law firm (*Same Claimant Law Firm*). The next two explanatory variables describe the differences in claim size between the two cases (*Absol. Claim Size Diff.*), and the differences in days between the times of the filing (*Absol. Filing Date Diff.*). We also include a dummy variable equal to one if customers in both cases had no legal representation (*Both Self-Represented*), and a series of dummy variables equal to one if both claims pertain to a specific controversy type (*Controversy Type Dummies*). To control for potential time trends, we also include a series of year indicators. Finally, we adopt specifications with case fixed effects to focus on variations in similarities across SOC pairs for a specific case. This method helps control for the effect of unobserved case characteristics on the similarity measure.

The results in Table 4 show that the content of two SOCs is more similar when they are prepared by the same attorney or law firm, when there is less difference in claim size, and when they are filed more closely in time. The economic magnitude of the attorney effect is the largest among the determinants. The coefficient estimate of *Same Claimant Attorney* is 0.287 after controlling for a full set of explanatory variables and case fixed effects (in Column 5). This indicates that SOCs prepared by the same attorney have average similarities that are 28.7 percent higher in absolute value than those by different attorneys. In comparison, the average pairwise similarity of the full sample is 11.4 percent.

Following the existing literature (Hanley and Hoberg, 2010), we also estimate the economic impact of the attorney effect using standard deviation units. Hanley and Hoberg (2010) find that IPOs with the same lead underwriter have overall document similarities that are 36.7 percent of one standard deviation higher than those with different lead underwriters. In our context, as

pairwise similarity has a standard deviation of 15.2 percent, SOCs written by the same attorney have similarities that are 1.88 times one standard deviation higher than SOCs written by different attorneys.[44]

6. The Empirical Evidence: Predicting Arbitration Outcomes

This section examines whether the text-based measures are useful in predicting arbitration outcomes as captured by the settlement decision and the total customer payout. We use logistic regressions to analyze the likelihood to settle, and OLS regressions to analyze total customer payout.[45] The three text-based variables capturing the information content of SOCs (i.e., *Length, Negative Tone, Avg. Similarity*) are the main predictors in our analyses. In the benchmark regressions, we include controls for case characteristics, controversy types, and year fixed effects. In some specifications, we also include controls for the firms named as respondents, hearing location, and arbitrator fixed effects.

6.1. The Likelihood to Settle

We test our second and third hypotheses by examining the relationship between case characteristics and the likelihood to settle. Table 5 reports the logistic regression results. We present the results using our textual measure as the stand-alone predictor in Columns 1,3, and 5; and the results when we include the other control variables in Columns 2,4, and 6. Consistent with Hypothesis 2, we find that stronger customer claims are associated with a higher likelihood to

---

[44] We acknowledge that the nature of documents in our sample may differ substantially from those in Hanley and Hoberg (2010). For example, we examine SOCs that arise from customer disputes in association with investments solely in Puerto Rico Municipal Bonds, whereas Hanley and Hoberg (2010) analyze prospectuses of all U.S. IPOs.

[45] For robustness, we also analyze the likelihood to settle with OLS regressions. These robustness results are reported in Appendix Table B2.

settle. We also find a similar relationship with claim similarity (Hypothesis 3). We find these results regardless of the specification.[46] Although document tone is negative and significant when it is the only regressor, it becomes insignificant when we include other controls.

See Table 5

The coefficient estimates of non-textual predictors are in line with past theory and empirical work. We find that large claims are more likely to go to award than settle. This is consistent with existing theory that parties are likely to have more divergent expectations concerning arbitration outcomes with larger claims (e.g., Priest and Clein, 1984; Posner, 1973). We also find that cases where customers represent themselves are more likely to go to award. Although the self-represented customers in the sample are associated with significantly smaller claims, these customers may be subject to larger errors in estimating outcomes due to their lack of expertise or experience.[47] As we present below, however, this lack of expertise or experience does not also result in significantly lower payouts on average.

See Table 6

Table 6 reports average changes in predicted probabilities of settlement when we vary the textual variables from the 25th to 75th percentile of the distribution while holding other determinants at the sample median. The results suggest that document similarity has the largest economic impact on settlement decisions. Across years, the likelihood of settlement increases 6.0 percentage points on average, from 91.3 to 97.3 percent, as document similarity increases from the

---

[46] In unreported analysis, we find similar results when we include indicators for firm, hearing location, and arbitrator fixed effects in the logistic regressions.

[47] Other factors, such as overconfidence of the self-represented customers, may also lead to more divergent expectations of outcomes between the parties and a lower likelihood of settlement (e.g., Korobkin and Guthrie, 1994; Kuhn, 2009).

25th to 75th percentile. Likewise, the likelihood of settlement increases 4.2 percentage points on average, from 91.7 to 95.9 percent, as document length increases from the 25th to 75th percentile.

6.2. Customer Payout

We next test the first hypothesis by examining the relationship between the text-based variables and total customer payout. Table 7 presents the results. Panel A reports the benchmark regressions results for all cases. Panels B and C report robustness results where we narrow the sample to those cases that settle or include additional fixed effects, respectively.

See Table 7

Consistent with Hypothesis 1, we find that stronger customer claims are associated with a higher payout-to-claim ratio and higher total payout. We also find similar relationships with claim similarity. There are multiple potential explanations for why similarity may relate to a higher payout. For example, attorneys may be capable of selecting cases that are more likely to result in a higher payout, leading to an attorney selection effect. They may also look for new cases that are similar to the ones they were previously able to settle. Finally, similar cases in our sample may share other traits that can be linked to a higher payout than those attributable to attorneys. Economically, a one standard deviation increase in document length (document similarity) is associated with a 2.6 percent (6.0 percent) increase in the payout-to-claim ratio. Given that the sample median total payout ratio is 28.5 percent, this corresponds to an increase in magnitude of 9.0 percent (21.6 percent) of the sample median level. A one standard deviation increase in document length (document similarity) is associated with a $25,740 ($19,860) increase in customer payout, a magnitude of 25.7 percent (19.9 percent) of the sample median ($100,000).

We find that document tone is negatively and significantly related to the payout-to-claim ratio and the total payout. A one standard deviation increase in negative document tone is associated

with a 6.4 percent ($48,960) decrease in the payout-to-claim ratio (total payout), a magnitude of 22.9 percent (49 percent) of the sample median. This is consistent with the notion that a customer often employs negative tone to add additional color in the SOC in an attempt to sway the arbitrators and other parties, potentially in place of additional facts or evidence which would support their claim.

Other control variables show that large claims tend to have lower payout-to-claim ratio but higher total payout. These results are consistent with existing work (e.g., Choi, Fisch, and Pritchard, 2010). We do not find evidence that self-represented customers are associated with significantly higher or lower payouts. Previous work suggests that legal representation can relate to higher awards (e.g., Choi, Fisch, and Pritchard, 2014). Other evidence suggests, however, that case complexity factors into the decision to self-represent (e.g., Swank, 2005). As we note above, the self-represented customers in the sample are associated with significantly smaller claim amounts, and approximately one-quarter of these claims are heard in arbitrations with simplified proceedings. Among all arbitrations in the sample, two percent are heard in arbitrations with simplified proceedings.

Overall, the results in Panel A suggest significant correlations exist between the textual measures and customer payout after controlling for case characteristics including claim size and self-representation. Existing studies (e.g., Egan, Matvos and Seru, 2020; Choi, Fisch, and Pritchard, 2010, 2014) document that hearing location, firm and arbitrator fixed effects can also significantly explain the variation of customer awards.[48] Our results in Panels B and C show that

[48] We note that our paper differs from these studies along many dimensions. For one, our sample is restricted to customer disputes concerning Puerto Rico municipal bonds and thus has less heterogeneity among hearing location, firm and arbitrator aspects. Furthermore, unlike existing work, our study accounts for the effects of case strength and similarity on customer awards.

our findings are robust to accounting for these fixed effects and restricting the sample to settled cases.

7. Robustness Tests

In this section, we discuss robustness tests. Results are provided in Tables B1, B2, and B3 in Appendix B.

7.1. Alternative Measures

First, we construct customer payout measures (i.e., *Payout-to-Claim Ratio_2 and Total Payout_2*) using an alternative assumption to combine the payout information from the Forum and CRD. We assume that the information is duplicative when more than one firm provides settlement information relating to the same case.[49] In these instances, we take the maximum reported settlement to estimate the settlement amount. In Panel A of Table B1, we present the results of regressions describing these alternative payout measures. The results are robust to these alternative customer payout measures.

As mentioned earlier, we analyze the longest SOC for each case to construct our key measures. It is likely that the last SOC filed for a case in our sample, potentially shorter, provides a better representation of the customer's case. To alleviate such a concern, we also construct alternative textual measures using the last SOC filed. Our main findings hold when we use these alternative measures as predictors. In Panel B of Table B1, we report the robustness of regressions describing the likelihood of settlement and the total payout-to-claim ratio.

7.2. Alternative Specifications

[49] As stated earlier, the majority of cases in our sample do not have multiple firms providing payout information associated with the same case.

Our evidence suggests that the information content of SOCs can predict arbitration decisions and outcomes. But such an effect may vary depending on whether a SOC relates to a self-represented claim. We thus include interaction terms between our textual variable and the variable indicating self-representation in our main regressions. We present the main results in Panel A of Table B2. The results in this table indicate that our findings are robust to this alternative specification.

We also acknowledge the possibility that the relationship between case strength, similarity and arbitration outcomes may not be linear. To account for a potential nonlinear effect, we include the squared terms of our textual variables. We present the key results in Panel B of Table B2. We find strong evidence about potential nonlinear effects between document length, similarity, and the likelihood to settle. Although in general the likelihood to settle increases with document length and similarity, for cases with very high values of document strength and similarity, the relationship becomes negative. The nonlinear effects are less persistent in regressions describing the total payout-to-claim ratio.

One potential concern is that parties may face constraints in the Forum process. For example, parties may have conflicts leading to difficulties in scheduling prehearing and hearing conferences or the Forum may face a limited supply of local arbitrators. These constraints may impact the decision to settle, the settlement amount, and the timing of the settlement.

To mitigate this concern, we construct two variables to proxy for the potential backlog of Puerto Rico cases: the number of open cases at a specific hearing location, and the number of days between the prehearing conference and the first scheduled hearing. The results are presented in Panel C of Table B2. The results show that parties are more likely to settle when they face more

constraints. The relationship between our textual variables and the likelihood to settle and total customer payout, however, remains unchanged.

Our main results also hold when we compute robust standard errors or standard errors clustered by year or by allegation type. We present the key results based on different types of standard errors in Panel D of Table B2.

7.3. Alternative Sample

Although most of our sample cases involve customer claims concerning violations of either fiduciary duty or suitability, 39 cases (or 1.2 percent) in our sample do not involve these allegations. As a final robustness check, to mitigate the potential impact of underlying differences in case characteristics, we exclude these 39 cases. We report the results in Table B3. The results are robust to the alternative sample.

8. Conclusion

Most customer disputes in securities arbitration settle rather than go to award. Despite this, there remains very little academic evidence pertaining to these settlements. In this paper, we examine the potential role of settlement as a means to resolve customer claims in securities arbitration. In particular, we investigate whether differences in case strength and similarity can explain arbitration parties' decision to settle or arbitrate a dispute.

Using a simple model rooted in the legal literature, we illustrate how parties in arbitration are more likely to settle when the parties anticipate a stronger customer claim or when the customer claim is more similar to previous claims. Empirically, we analyze the SOCs filed in the Forum and relate SOC characteristics to arbitration outcomes. We narrow the sample to cases concerning investments in Puerto Rico municipal bonds only. This provides us with a relatively clean setting

by holding constant such factors as the type of the financial product, macro conditions, industry party characteristics, and the violations customers assert.

We first develop textual measures and show that claimant attorneys have a large influence on the content of these documents. We then link these measures to the likelihood of settlement and the total payout. Consistent with the simple model, we find that strong claims and claims that are similar to previous claims are more likely to settle. These claims are also associated with higher customer payout. Our results are robust to alternative textual and payout measures; alternative regression specifications accounting for nonlinear effects, robust standard errors, and clustered standard errors; and an alternative sample consisting of only those customer allegations related to violations of fiduciary duty or suitability.

Overall, our paper provides the first empirical analysis of settlement decisions in securities arbitration. The findings suggest that cases that settle may share different characteristics from those cases that instead result in an award. As all prior studies on securities arbitration concern cases that result in an award, our research suggests that inferences about arbitration outcomes made from just those cases may be subject to a severe selection bias.

Reference

Amicis, C D, S Falconieri, M Tastan, Forthcoming, “Sentiment analysis and gender differences in earnings conference calls”, *Journal of Corporate Finance.*

Baker, Scott R., Nicholas Bloom, Steven J. Davis, 2016, “Measuring economic policy uncertainty”, *Quarterly Journal of Economics* 131 (4): 1593-1636.

Bebchuck, Lucian Arye, 1984, “Litigation and settlement under imperfect information”, *RAND Journal of Economics* 15: 404-15.

Bloom, David E., 1986, "Empirical models of arbitrator behavior under conventional arbitration," *Review of Economics and Statistics* 68: 578-85.

Bloom, David E., and Christopher L. Cavanagh, 1986, “An analysis of the selection of arbitrators”, *The American Economic Review* 76(3):408–422.

Choi, Stephen J., and T Eisenberg, 2010, “Punitive damages in securities arbitration: an empirical study”, *Journal of Legal Studies* 39 (2): 497-546.

Choi, Stephen J., Jill E. Fisch, and A. C. Pritchard, 2010, “Attorneys as arbitrators”, *Journal of Legal Studies* 39: 109-57.

Choi, Stephen J., Jill E. Fisch, and A. C. Pritchard, 2014, “The influence of arbitrator background and representation on arbitration outcomes”, *Virginia Law and Business Review* 9: 43-90.

Correa, R., K. Garud, J. Londono, and N. Mislang, 2021, “Sentiment in central banks’ financial stability reports”, *Review of Finance* 25, 85-120. Crawford, Vincent P., 1979, “On compulsory-arbitration schemes”, *Journal of Political Economy* 87: 131-59.

Dickson, David L., 2005, “Bargaining outcomes with double-offer arbitration”, *Experimental Economics* 8(2): 145-66.

Egan, Mark, Gregor Matvos, and Amit Seru, 2020, "Arbitration with uninformed consumers", *Harvard Business School Working Paper.*

Faber, Henry S., and Max H. Bazerman, 1989, "Divergent expectations as a cause of disagreement in bargaining: Evidence from a comparison of arbitration schemes," *Quarterly Journal of Economics* 104(1):99-120.

Farber, Henry S., Margaret A. Neale, and Max H. Bazerman, 1990, "The role of arbitration costs and risk aversion in dispute outcomes", *Industrial Relations* (29) 3: 361-84.

Farmer, Amy, and Paul Pecorino, 1998, "Bargaining with informative offers: An analysis of final-offer arbitration", *Journal of Legal Studies* 27(2): 415-32.

FINRA, 2018, "Discussion Paper – FINRA Perspectives on Customer Recovery," https://www.finra.org/sites/default/files/finra_perspectives_on_customer_recovery.pdf.

Fresard, Laurent, Gerald Hoberg, and Gordon Phillips, 2020, "Innovation activities and integration through vertical acquisitions ", *Review of Financial Studies* 33 (7): 2937-76.

Gould, John P., 1973, "The Economics of Legal Conflicts", *The Journal of Legal Studies* 2 (2): 279-300.

Hancock, Jeffrey T., Lauren E. Curry, Saurabh Goorha, and Michael Woodworth, 2007, "On lying and being lied to: A linguistic analysis of deception in computer-mediated communication, *"Discourse Processes* 45 (1), 1-23.

Hanley, Kathleen, and Gerard Hoberg, 2010, "The information content of IPO prospectus", *Review of Financial Studies* 23 (7): 2821–64.

Hanley, Kathleen, and Gerard Hoberg, 2012, "Litigation risk, strategic disclosure and the underpricing of initial public offerings", *Journal of Financial Economics* 103 (2): 235-54.

Hanley, Kathleen, and Gerald Hoberg, 2019, "Dynamic interpretation of emerging risks in the financial sector", *Review of Financial Studies* 32 (12): 4543–4603.

Helland, E., Klerman, D., and Lee, Y. H. A., 2018, "Maybe there is no bias in the selection of disputes for litigation", *Journal of Institutional and Theoretical Economics* 174: 143-170

Hoberg, Gerard, and Gordon Phillips, 2010, "Product market synergies and competition in mergers and acquisitions: A text-based analysis," *Review of Financial Studies* 23 (10): 3773-811.

Hoberg, Gerard, and Gordon Phillips, 2016, "Text-based network industries and endogenous product differentiation", *Journal of Political Economy* 124 (5): 1423-65.

Honigsberg, Colleen, and Matthew Jacob, 2021, "Deleting misconduct: The expungement of BrokerCheck records", *Journal of Financial Economics* 139: 800-831.

Huang, X, SH Teoh, Y Zhang, 2014, " Tone management", The Accounting Review 89:1083-113.

Keith Hylton, 1993, "Asymmetric information and the selection of disputes for litigation", *Journal of Legal Studies* 22: 187-210.

Kessler, Daniel, Thomas Meites, and Geoffrey Miller, 1996, "Explaining deviations from the Fifty-Percent Rule: A multimodal approach to the selection of cases for litigation", *Journal of Legal Studies* 25:233-59.

Klerman, Daniel, 2012, "The selection of 13th-century disputes for litigation", *Journal of Empirical Legal Studies* 9: 320-46.

Kondo, Jiro E., 2007, "The self-regulation of enforcement: Evidence from investor-broker disputes at the NASD", *MIT Dissertation.*

Korobkin, Russell, and Chris Guthrie, 1994, "Psychological barriers to litigation settlement: An experimental approach", *Michigan Law Review* 93: 107-192.

Korobkin, Russell, and TS Ulen, 2000, "Law and behavioral science: Removing the rationality assumption from law and economics", California Law Review 88 (4):1051-63.

Kozora Matthew, 2017, "Security recommendations and the liabilities of broker-dealers", *Journal of Law, Finance and Accounting* 2: 385-428.

Kuhn, Michael A., 2009, "To settle or not to settle: A review of the Literature on arbitration in the laboratory," *Manuscript, University of California, San Diego*.

Landes, William M., 1971, "An Economic Analysis of the Courts", *The Journal of Law & Economics* 14 (1): 61-107.

Lee, Yoon-Ho Alex, and Daniel Klerman, 2016, "The Priest-Klein hypotheses: proofs and generality", *International Review of Law and Economics* 48: 59-76.

Loughran, Tim, and Bill McDonald, 2014, "Regulation and financial disclosure: The impact of plain English", *Journal of Regulatory Economics* 45, 94-113.

Loughran, Tim, and Bill McDonald, 2011, "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks", *Journal of Finance* 66, 35–65.

Loughran, Tim, and Bill McDonald, 2020, "Textual analysis in finance", *Annual Review of Financial Economics* 12, 357–375.

Merkl-Davies, D., & Brennan, N., 2007, "Discretionary disclosure strategies in corporate narratives: Incremental information or impression management?" *Journal of Accounting Literature* 26,116-196.

Pecorino, Paul and Mark van Boening, 2001, "Bargaining and information: An empirical analysis of a multistage arbitration game", *Journal of Labor Economics* 19(4): 922-48.

Posner, Richard A., 1973, "An economics approach to legal procedure and judicial administration", *Journal of Legal Studies* 2: 399-458.

Priest, George, and Benjamin Klein, 1984, "The selection of disputes for litigation," *Journal of Legal Studies* 13, 1-55.

Rachlinkski, 1996, "Gains, losses, and the psychology of litigation", *Southern California Law Review* 70,113-85.

Reinganum, Jennifer E., and Louis L. Wilde, 1986, "Settlement, Litigation, and the Allocation of Litigation Costs", *The RAND Journal of Economics* 17(4):557-566.

Shavell, Steven, 1996, "Any frequency of plaintiff victory at trail is possible", *Journal of Legal Studies* 25: 493-501.

Schlenker, B. R., 1980, "Impression management: The self-concept, social identity, and interpersonal relations", Monterey/California: Brooks/Cole.

Studdert, David M., and Michelle M. Mello, 2007, "When tort resolutions are "wrong": predictors of discordant outcomes in medical malpractice litigation", *Journal of Legal Studies* 36:547-78.

Swank, Drew A., 2005, "The Pro Se Phenomenon", *Brigham Young University Journal of Public Law* 19(2):373-86.

Toma, Catalina L. and Jeffrey T. Hancock, 2012, "What lies beneath: The linguistic traces of deception in online dating profiles, *Journal of Communication* 62 (1), 78-97.

You, Haifeng and XJ Zhang, 2009, "Financial reporting complexity and investor underreaction to 10-K information", *Review of Accounting Studies* 14:559-586.

Figure 1: Timeline of Securities Arbitration

This figure depicts the timeline of securities arbitration.

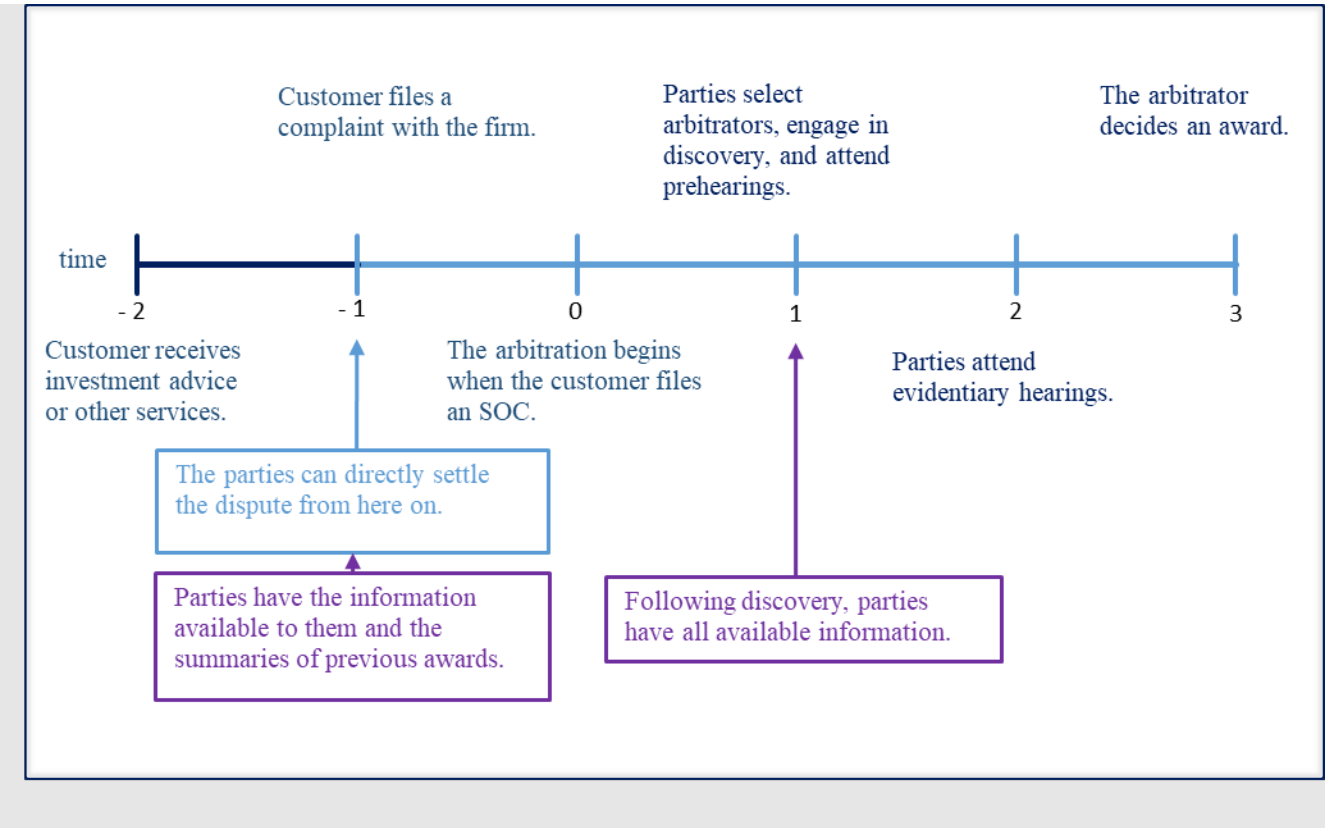

Figure 2: Top 100 Meaningful Words

This figure depicts the top 100 meaningful words used in the collection of SOCs in our sample cases. The bigger and bolder a specific word is, the more often it is mentioned in the texts. The sample consists of 3,325 customer claims filed and closed in FINRA's Forum between January 2014 and September 2020 pertaining to investments in Puerto Rico municipal bonds.

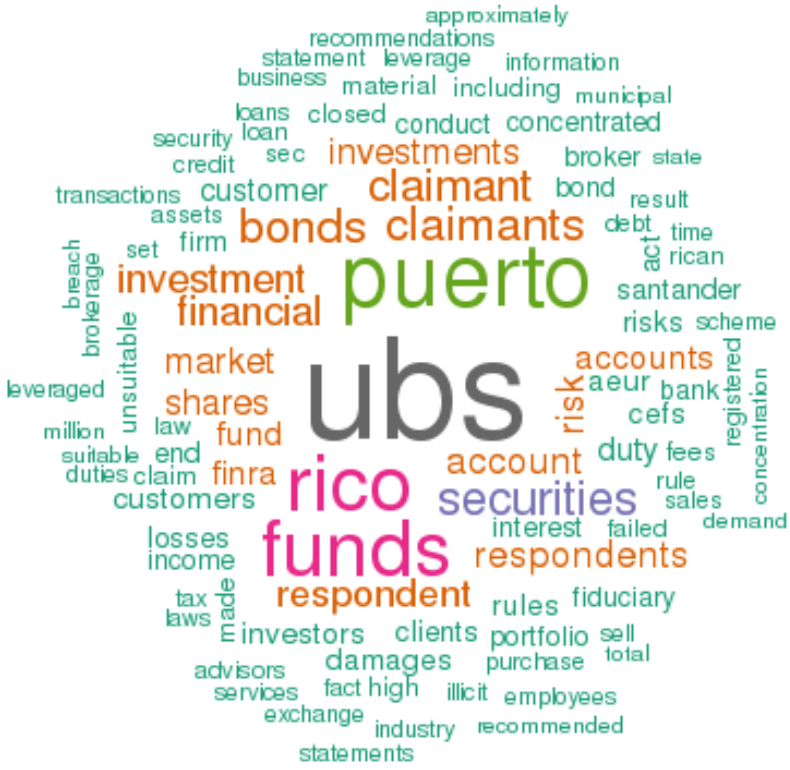

Table 1: Characteristics of the Textual Variables

This table reports the distribution for text-based variables. Panel A reports the number of observations, sample minimum, 25th percentile, median, 75th percentile, maximum, mean, and standard deviation for each variable, respectively. Panel B compares above-the-mean claim size group (Group 1) versus below-the-mean claim size group (Group 2), and self-represented claims (Group 1) versus claims with legal representation (Group 2). Reported are the average values by group and differences in averages. Coefficients marked with *, **, and *** are significant at the 0.1, 0.05, and 0.01 levels, respectively. The sample consists of 3,235 customer claims filed and closed in the FINRA Forum between January 2014 and September 2020 pertaining to investments in Puerto Rico municipal bonds. Variables are defined in Appendix A.

Panel A: Descriptive Statistics

	Length	Negative Tone	Similarity	
			Pairwise	Avg
	(1)	(2)	(3)	(4)
Obs	3,235	3,235	5,230,995	3,235
Min	91	0.7%	0.0%	0.1%
P25	2,204	7.0%	2.1%	4.5%
P50	3,928	7.8%	5.6%	8.3%
P75	5,632	8.8%	14.9%	13.4%
Max	15,070	21.8%	100.0%	22.6%
Mean	4,215	8.0%	11.4%	8.0%
Stdev	2,574	1.7%	15.2%	5.5%

Panel B: Distribution by Case Characteristics

	Length			Negative Tone			Avg Similarity		
	Group 1	Group 2	Diff (1)-(2)	Group 1	Group 2	Diff (1)-(2)	Group 1	Group 2	Diff (1)-(2)
Case Characteristics	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Claim Size (10 mil$)	5,632	3,928	1,704***	7.80%	8.80%	-1%**	8.30%	13.40%	-5.1%**
Whether Self-Represented	2,204	5,632	-3,428***	8.80%	7.00%	1.8%***	4.50%	13.40%	-8.9%***

Table 2: Sample Description

This table describes the sample. Panels A, B, C, and D report the number of observations by year, firm, hearing location, and controversy types, respectively. The sample consists of 3,235 customer claims filed and closed in the FINRA Forum between January 2014 and September 2020 pertaining to investments in Puerto Rico municipal bonds.

Panel A: Distribution by Year

	2014	2015	2016	2017	2018	2019	2020	Sum
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
No of Cases	678	589	545	619	611	177	16	3,235
% of All Cases	20.96%	18.21%	16.85%	19.13%	18.89%	5.47%	0.49%	100.00%

Panel B: Distribution by Major Firms

	UBS	Santander Securities	Popular Securities	Merrill Lynch	Oriental Financial Group	Sum
	(1)	(2)	(3)	(4)	(5)	(6)
No of Cases	1955	470	160	112	102	2,799
% of All Cases	60.43%	14.53%	4.95%	3.46%	3.15%	86.52%

Panel C: Distribution by Major Hearing Locations

	San Juan, Puerto Rico	New York, NY	Boca Raton, FL	Atlanta, GA	Miami, FL	Sum
	(1)	(2)	(3)	(4)	(5)	(6)
No of Cases	2,218	270	134	120	92	2,834
% of All Cases	68.56%	8.35%	4.14%	3.71%	2.84%	87.60%

Panel D: Distribution by Major Controversy Types

	Breach of Fiduciary Duty or Negligence	Failure to Supervise	Misrepresentation or Omission of Facts	Breach of Contract	Suitability	Sum
	(1)	(2)	(3)	(4)	(5)	(6)
No of Cases	3,146	2,717	2,706	2,673	2,578	3,208
% of All Case	97.25%	83.99%	83.65%	82.63%	79.69%	99.17%

Table 3: Summary Statistics

This table reports the summary statistics for non-textual variables. Columns 1-8 report the number of observations, sample minimum, 25th percentile, median, 75th percentile, maximum, mean, and standard deviation for each variable, respectively. The sample consists of 3,235 customer claims filed and closed in the FINRA Forum between January 2014 and September 2020 pertaining to investments in Puerto Rico municipal bonds. Variables are defined in Appendix A.

	Obs	Min	P25	P50	P75	Max	Mean	Stdev
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Case Characteristics								
Whether Self-Represented	3,235	0.00	0.00	0.00	0.00	1.00	0.01	0.11
Claim Size (10 mil$)	3,235	0.00	0.02	0.04	0.10	225	0.20	3.98
Whether Unspecified Damages	3,235	0.00	0.00	0.00	0.00	1.00	0.08	0.27
Panel B: Arbitration Outcomes								
Whether Going through Hearings	3,235	0.00	0.00	0.00	0.00	1.00	0.03	0.18
Whether Decided on Paper	3,235	0.00	0.00	0.00	0.00	1.00	0.01	0.07
Whether Settled	3,235	0.00	1.00	1.00	1.00	1.00	0.92	0.28
Whether Withdrawn	3,235	0.00	0.00	0.00	0.00	1.00	0.03	0.18
Total Payout (10 mil$)	3,051	0.00	0.01	0.01	0.03	3.75	0.04	0.14
Payout-to-Claim Ratio	3,051	0.00	0.15	0.29	0.48	43.01	0.44	1.18
Panel C: Claimant Legal Representation								
	No	**Cases Per Attorney (Law Firm)**						
		Min	**P25**	**P50**	**P75**	**Max**	**Mean**	**Stdev**
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Attorneys	412	1	1	1	2	378	9	32
Repeated Attorneys	143	2	2	5	19	378	24	50
Repeated Law Firms	155	2	2	4	17	519	22	60

Table 4: Sources of SOC Content

This table reports OLS regression results on the sources of SOC content. The dependent variable is document similarity of two SOCs. One observation is one pair of SOCs for any two different cases. The sample consists of 3,235 customer claims filed and closed in the FINRA Forum between January 2014 and September 2020 pertaining to investments in Puerto Rico municipal bonds. We include common controversy type indicators and year indicators in models (2), (3), (5), (6), and case fixed effects in models (4)-(6). Coefficients marked with *, **, and *** are significant at the 0.1, 0.05, and 0.01 levels, respectively. Robust standard errors clustered by year are reported in parenthesis. Variables are defined in Appendix A.

	Pairwise Similarity					
	(1)	(2)	(3)	(4)	(5)	(6)
Same Claimant Attorney	0.307***	0.294***		0.298***	0.287***	
	(0.026)	(0.024)		(0.024)	(0.022)	
Same Claimant Law Firm			0.254***			0.268***
			(0.014)			(0.014)
Absol. Claim Size Diff.		-0.001***	-0.001***		-0.0004***	-0.0004***
		(0.000)	(0.000)		(0.000)	(0.000)
Absol. Filing Date Diff.		-0.00002**	-0.00002**		-0.00001**	-0.00001***
		(0.000)	(0.000)		(0.000)	(0.000)
Both Self-Represented		-0.034***	-0.032***		0.011**	0.010*
		(0.006)	(0.005)		(0.005)	(0.005)
Controversy Types Dummies	N	Y	Y	N	Y	Y
Year Dummies	N	Y	Y	N	Y	Y
Case Dummies	N	N	N	Y	Y	Y
No. of Obs.	5,230,995	5,230,995	5,230,995	5,230,995	5,230,995	5,230,995
R-Square	0.128	0.167	0.169	0.136	0.165	0.181

Table 5: Likelihood to Settle

This table examines the determinants of likelihood to settle. The dependent variable is an indicator that equals one if a case was settled and zero otherwise. The sample consists of 3,235 customer claims filed and closed in the FINRA Forum between January 2014 and September 2020 pertaining to investments in Puerto Rico municipal bonds. Logistic regression results are reported with standard errors in parentheses. Coefficients marked with *, **, and *** are significant at the 0.1, 0.05, and 0.01 levels, respectively. Variables are defined in Appendix A.

	Whether Settled					
	(1)	(2)	(3)	(4)	(5)	(6)
Avg. Similarity	14.114***	11.433***				
	(1.508)	(1.545)				
Length			0.0002***	0.0002***		
			(0.000)	(0.000)		
Negative Tone					-8.336**	-2.218
					(3.841)	(3.886)
Claim Size		-0.386***		-0.427***		-0.379***
		(0.107)		(0.113)		(0.105)
Whether Self-Represented		-1.506***		-1.734***		-1.866***
		(0.355)		(0.351)		(0.354)
Controv. Type Dummies	N	Y	N	Y	N	Y
Year Dummies	N	Y	N	Y	N	Y
Obs.	3,235	3,222	3,235	3,222	3,235	3,222
Pseudo R-Square	0.137	0.197	0.103	0.183	0.068	0.156

Table 6: Economic Significance of Textual Variables

This table examines average changes in predicted likelihood to settle when *Avg. Similarity* (Panel A), *Length* (Panel B) and *Negative Tone* (Panel C) vary from 25th percentile to 75th percentile based on the logistic models in Columns 2, 4, and 6 of Table 5, respectively. To compute change in predicted probability, we hold other determinants at the sample median (except holding *Whether Self-Represented* at the value of zero). The sample consists of 3,235 customer claims filed and closed in the FINRA Forum between January 2014 and December 2018 pertaining to investments in Puerto Rico municipal bonds. Variables are defined in Appendix A.

Panel A: Change in Predicted Probability When Avg. Similarity Varies

	Avg. Similarity		
	P25	P75	Diff (1)-(3)
Year	(1)	(2)	(3)
2014	91.9%	97.7%	5.8%
2015	90.2%	96.5%	6.3%
2016	89.0%	95.7%	6.7%
2017	91.8%	97.9%	6.1%
2018	93.6%	98.5%	4.9%
Mean	91.3%	97.3%	6.0%

Panel B: Change in Predicted Probability When Length Varies

	Length		
	P25	P75	Diff (1)-(3)
Year	(1)	(2)	(3)
2014	92.5%	96.4%	3.9%
2015	91.2%	95.2%	4.0%
2016	88.5%	93.7%	5.2%
2017	92.1%	96.4%	4.3%
2018	94.3%	97.6%	3.3%
Mean	91.7%	95.9%	4.1%

Panel C: Change in Predicted Probability When Negative Tone Varies

	Negative Tone		
	P25	P75	Diff (1)-(3)
Year	(1)	(2)	(3)
2014	92.3%	92.0%	-0.3%
2015	92.7%	92.4%	-0.3%
2016	89.4%	89.0%	-0.4%
2017	92.7%	92.4%	-0.3%
2018	94.4%	94.2%	-0.2%
Mean	92.3%	92.0%	-0.3%

Table 7: Determinants of Customer Payout

This table examines the determinants of customer payout. Panel A reports the benchmark results. Panel B reports the subsample results for settled cases. Panel C reports results after controlling for additional fixed effects. In Panels A and B, the dependent variable is *Payout-to-Claim Ratio* in Columns 1-3 and *Total Payout* in Columns 4-6, respectively. The sample consists of 3,235 customer claims filed and closed in the FINRA Forum between January 2014 and September 2020 pertaining to investments in Puerto Rico municipal bonds. OLS regression results are reported with standard errors in parentheses. Coefficients marked with *, **, and *** are significant at the 0.1, 0.05, and 0.01 levels, respectively. Variables are defined in Appendix A.

Panel A: Benchmark Results

	Payout-to-Claim Ratio			Total Payout		
	(1)	(2)	(3)	(4)	(5)	(6)
Avg. Similarity	1.089***			0.036***		
	(0.419)			(0.004)		
Length		0.00001**			0.000002***	
		(0.000)			(0.000)	
Negative Tone			-3.742***			-0.377***
			(1.290)			(0.097)
Claim Size	-0.113**	-0.118**	-0.109**	0.112***	0.111***	0.112***
	(0.050)	(0.050)	(0.050)	(0.005)	(0.005)	(0.005)
Whether Self-Represented	-0.084	-0.120	-0.166	-0.016	-0.015	-0.020
	(0.212)	(0.211)	(0.211)	(0.023)	(0.022)	(0.022)
Controv. Type Dummies	Y	Y	Y	Y	Y	Y
Year Dummies	Y	Y	Y	Y	Y	Y
Obs.	3,040	3,040	3,040	3,040	3,040	3,040
R-Square	0.013	0.011	0.013	0.141	0.142	0.143

Panel B: Subsample Results for Settled Cases

	Payout-to-Claim Ratio			Total Payout		
	(1)	(2)	(3)	(4)	(5)	(6)
Avg. Similarity	1.211***			0.026**		
	(0.432)			(0.012)		
Length		0.00001**			0.000001**	
		(0.000)			(0.000)	
Negative Tone			-4.079***			-0.288**
			(1.350)			(0.122)
Claim Size	-0.156**	-0.165**	-0.154**	0.141***	0.139***	0.141***
	(0.066)	(0.067)	(0.066)	(0.006)	(0.006)	(0.006)
Whether Self-Represented	-0.073	-0.111	-0.134	-0.016	-0.016	-0.018
	(0.274)	(0.274)	(0.273)	(0.025)	(0.025)	(0.025)
Controv. Type Dummies	Y	Y	Y	Y	Y	Y
Year Dummies	Y	Y	Y	Y	Y	Y
Obs.	2,816	2,816	2,816	2,816	2,816	2,816
R-Square	0.013	0.011	0.013	0.176	0.176	0.177

Panel C: Controlling for Additional Fixed Effects

Row	Avg. Similarity	Length	Negative Tone	Hearing Location and Firm Dummies	Arbitrator Dummies	Obs (R-square)
			Dependent Variable: Payout-to-Claim Ratio			
(1)	0.631**			Yes	No	3,040
	(0.224)					(0.011)
(2)	0.852***			Yes	Yes	8,116
	(0.308)					(0.012)
(3)		0.000004**		Yes	No	3,040
		(0.000)				(0.011)
(4)		0.000002*		Yes	Yes	8,116
		(0.000)				(0.011)
(5)			-3.134***	Yes	No	3,040
			(1.198)			(0.013)
(6)			-4.671***	Yes	Yes	8,116
			(0.929)			(0.015)
			Dependent Variable: Total Payout			
(7)	0.027***			Yes	No	3,040
	(0.002)					(0.141)
(8)	0.029***			Yes	Yes	8,116
	(0.000)					(0.130)
(9)		0.0000004*		Yes	No	3,040
		(0.000)				(0.141)
(10)		0.0000001*		Yes	Yes	8,116
		(0.000)				(0.130)
(11)			-0.262**	Yes	No	3,040
			(0.103)			(0.142)
(12)			-0.037***	Yes	Yes	8,116
			(0.000)			(0.133)

Appendix A. Variable Definitions

Variable	Definition
Arbitration Outcome Variables	
Whether Going through Hearings	An indicator that equals one if a case was closed by going through a hearing, and zero otherwise.
Whether Decided on Paper	An indicator that equals one if a case was decided by an arbitrator based on the parties' pleadings and other written submissions, and zero otherwise.
Whether Settled	An indicator that equals one if a case was closed by settlement, and zero otherwise.
Whether Withdrawn	An indicator that equals one if a case was withdrawn, and zero otherwise.
Award (10 mil$)	The amount of award in 10 million dollars.
Total Payout (10 mil$)	Total payout to customers including award and settlement amount in 10 million dollars.
Payout-to-Claim Ratio	The ratio of *Total Payout* divided by *Claim Size*.
Textual Variables	
Length	Total number of meaningful words in the main texts of an SOC (exluding exhibits).
Negative Tone	Percentage of meaningful words that are negative words (Loughran and McDonald, 2011).
Pairwise Similarity	Cosine similarity of two-SOC word vectors. See formla in Hanley and Hoberg (2010).
Avg. Similarity	The average of pairwise similarity with other previously-filed SOCs for a specific SOC.
Case Characteristics	
Whether Self-Represented	An indicator that equals one if the customer in a specific case had no legal representation, and zero otherwise.
Claim Size (10 mil$)	The amount of damage claimed by the customer in 10 millin dollars. If the damage amount was not specified, we set the claim size at the sample median.
Whether Unspecified Damages	An indicator that equals one if the damage amount was not specified in a claim, and zero otherwise.
Claimant Legal Representation	
Attorneys	The attorneys that represented customers in our sample.
Repeated Attorneys	Attorneys that represented more than two customer cases in our sample.
Repeated Law Firms	Law firms that represented more than two customer cases in our sample.
Other Controls	
Same Claimant Attorneys	An indicator that equals one if two SOCs were prepared by the same attorney, and zero otherwise.
Same Claimant Law Firms	An indicator that equals one if two SOCs were prepared by the same law firm, and zero otherwise.
Absol. Claim Size Diff.	The absolute value of difference in claim size between two cases.
Absol. Filing Date Diff.	The absolute value of difference in filing dates (in days) between two cases.
Both Self-Represented	An indicator that equals one if the customers in both cases had no legal representation, and zero otherwise.
Year Dummies	A series of year indicators (year 2014/2015/2016/2017/2018/2019) that equal one for observations in a specific year (year 2014/2015/2016/2017/2018/2019), and zero otherwise.
Controversy Type Dummies	A series of controversy type indicators that equal one if a claim pertains to a specific controversy type, and zero otherwise. We include the following six controversy type dummies: Breach of Fiduciary Duty or Neglience; Failure to Supervise; Misrepresentation or Omission of Facts; Suitability; Breach of Contract; Unauthorized Trading or Churning or Manipulation.
Case Dummies	A series of case indicators that equal one for observations related to a specific case, and zero otherwise.
Firm Dummies	A series of firm indicators that equal one for claims against a specific firm, and zero otherwise.
Hearing Location Dummies	A series of location indicators that equal one for claims related to a specific hearing location, and zero otherwise.
Arbitrator Dummies	A series of arbitrator indicators that equals one if a specific arbitrator was on the panel of a specific case, and zero otherwise.
Days between Prehearing and 1st Scheduled Hearing	Number of days between the initial prehearing conference and first scheduled evidentiary hearing.
No. of Local Open Cases	Number of open Puerto-Rico cases at a specific hearing location.

Appendix B. Robustness Checks

Table B1: Alternative Measures

This table reports estimation results using alternative measures of customer payout (Panel A) and textual variables (Panel B). Panel A reports the estimation results from Table 7 based on alternative estimates for settlement amount (i.e., two alternative measures—*Payout-to-Claim Ratio_2, Total Payout_2*). Panel B reports estimation results from Columns 2, 4, 6 of Table 5 and Columns 4-6 of Table 7 using alternative textual measures based on the latest SOCs for each case (i.e., *Avg Similarlity_2*, *Length_2*, and *Negative Tone_2*). The sample consists of 3,235 customer claims filed and closed in the FINRA Forum between January 2014 and September 2020 pertaining to investments in Puerto Rico municipal bonds. OLS (Logistic) regression results are reported with standard errors in parentheses in Columns 1-6 in Panel A and Columns 4-6 in Panel B (Columns 1-3 in Panel B). Other controls include *Claim Size*, *Whether Self-Represented*, and Controversy type and year dummies. Coefficients marked with *, **, and *** are significant at the 0.1, 0.05, and 0.01 levels, respectively. Variables are defined in Appendix A.

Panel A: Alternative Payout Measures

	Payout-to-Claim Ratio_2			**Total Payout_2**		
	(1)	(2)	(3)	(4)	(5)	(6)
Avg. Similarity	0.819**			0.0095**		
	(0.364)			(0.004)		
Length		0.000005**			0.000002***	
		(0.000)			(0.000)	
Negative Tone			-3.175***			-0.304***
			(1.121)			(0.097)
Other Controls	Y	Y	Y	Y	Y	Y
Obs.	3,040	3,040	3,040	3,040	3,040	3,040
R-Square	0.010	0.009	0.011	0.136	0.136	0.139

Panel B: Alternative Textual Measures

	Whether Settled			Payout-to-Claim Ratio		
	(1)	(2)	(3)	(4)	(5)	(6)
Avg. Similarity_2	11.433***			0.978**		
	(1.823)			(0.408)		
Length_2		0.0002***			0.00002*	
		(0.000)			(0.000)	
Negative Tone_2			-11.673***			-2.357**
			(3.791)			(1.081)
Other Controls	Y	Y	Y	Y	Y	Y
Obs.	2,670	2,670	2,670	2,569	2,569	2,569
Pseudo R-Square(R-Square)	0.187	0.173	0.146	0.013	0.011	0.012

Table B2: Alternative Specifications

This table reports estimation results from Columns 2, 4, 6 of Table 5 and Columns 4-6 of Table 7 using alternative specifications. Panel A reports OLS estimation results with additional interaction terms between the textual measures and *Whether Self-Represented*. Panel B reports OLS regression results including squared textual variables. Panel C reports logistic and OLS regressions after controlling for a potential backlog of Puerto Rico cases. Panel D reports logistic and OLS regression results based on robust and clustered standard errors. The sample consists of 3,235 customer claims filed and closed in the FINRA Forum between January 2014 and September 2020 pertaining to investments in Puerto Rico municipal bonds. Other controls include *Claim Size*, *Whether Self-Represented*, and Controversy type and year dummies. In Panels A, B and C, coefficients marked with *, **, and *** are significant at the 0.1, 0.05, and 0.01 levels, respectively, with standard errors reported in parentheses. In Panel D, conventional standard errors, robust standard errors, standard errors clustered by year, and standard errors clustered by controversy types are reported in parentheses. Standard errors marked with *, **, and *** are associated with coefficient estimates significant at the 0.1, 0.05, and 0.01 levels. Variables are defined in Appendix A.

Panel A: Adding Interaction Terms

	Whether Settled (OLS)						Payout-to-Claim Ratio (OLS)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Avg. Similarity	0.0219***			0.0219***			0.785**		
	(0.025)			(0.025)			(0.329)		
Avg. Similarity				0.272			-2.457		
* Whether Self-Represented				(0.431)			(5.608)		
Length		0.000001*			0.000001*			0.000004*	
		(0.000)			(0.000)			(0.000)	
Length					-0.010			0.000	
* Whether Self-Represented					(0.017)			(0.000)	
Negative Tone			-0.161**			-0.161**			-2.678093**
			(0.080)			(0.080)			(1.044)
Negative Tone						-0.030			1.112
* Whether Self-Represented						(0.026)			(4.753)
Other Controls	Y	Y	Y	Y	Y	Y	Y	Y	Y
Obs.	3,235	3,235	3,235	3,235	3,235	3,235	3,235	3,235	3,235
R-Square	0.197	0.183	0.156	0.197	0.183	0.156	0.013	0.011	0.012

Panel B: Adding Squared Textual Variables

	Whether Settled (OLS)			Payout-to-Claim Ratio (OLS)		
	(1)	(2)	(3)	(4)	(5)	(6)
Avg. Similarity	1.895***			2.397*		
	(0.356)			(1.675)		
Length		0.0002***			0.00001**	
		(0.000)			(0.000)	
Negative Tone			3.701***			-5.336*
			(1.325)			(2.690)
Avg. Similarity_Squared	-5.841***			-6.139		
	(1.658)			(7.744)		
Length_Squared		-0.00000001***			-0.00000001***	
		(0.000)			(0.000)	
Negative Tone_Squared			-21.858***			8.722
			(7.470)			(36.164)
Other Controls	Y	Y	Y	Y	Y	Y
Obs.	3,222	3,222	3,222	3,040	3,040	3,040
R-Square	0.122	0.116	0.105	0.013	0.016	0.013

Panel C: Controlling for Backlogs of Puerto Rico Cases

	Whether Settled (Logit)			Payout-to-Claim Ratio (OLS)		
	(1)	(2)	(3)	(4)	(5)	(6)
Avg. Similarity	10.805***			1.679***		
	(2.125)			(0.479)		
Length		0.0001***			0.00002**	
		(0.000)			(0.000)	
Negative Tone			1.911			-4.3328***
			(4.425)			(1.383)
Days between Prehearing and	0.002***	0.002***	0.003***	-0.0004***	-0.0004***	-0.0003***
1st Scheduled Hearing	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)
No. of Local Open Cases	0.0003*	0.0003*	0.0005**	0.00009	0.00008	0.0001
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Other Controls	Y	Y	Y	Y	Y	Y
Obs.	3,046	3,047	3,047	2,893	2,893	2,893
Pseudo R-Square/R-Square	0.363	0.352	0.348	0.015	0.013	0.015

Panel D: Results with Robust and Clustered Standard Errors

	Whether Settled (Logit)			Payout-to-Claim Ratio (OLS)		
	(1)	(2)	(3)	(4)	(5)	(6)
Avg. Similarity	11.433			1.089		
Conventional S.E.	(1.545)***			(0.419)***		
Robust S.E.	(1.649)***			(0.511)**		
S.E. Clustered by Year	(1.914)***			(0.608)*		
S.E. Clustered by Controversy Type	(2.558)***			(0.471)***		
Length		0.0002			0.00001	
Conventional S.E.		(0.000)***			(0.000)**	
Robust S.E.		(0.000)***			(0.000)***	
S.E. Clustered by Year		(0.000)***			(0.000)*	
S.E. Clustered by Controversy Type		(0.000)***			(0.000)***	
Negative Tone			-2.218			-3.742
Conventional S.E.			(3.886)			(1.290)***
Robust S.E.			(4.372)			(0.836)***
S.E. Clustered by Year			(2.090)			(0.918)***
S.E. Clustered by Controversy Type			(5.149)			(0.824)***

Table B3: Alternative Sample

This table reports estimation results from Columns 2, 4, 6 of Table 5 and Columns 4-6 of Table 7 based on an alternative sample. The alternative sample consists of 3,196 customer claims filed and closed in the FINRA Forum between January 2014 and September 2020 concerning violations of fiduciary duty or suitability regarding investments in Puerto Rico municipal bonds. Other controls include *Claim Size*, *Whether Self-Represented*, and Controversy type and year dummies. Coefficients marked with *, **, and *** are significant at the 0.1, 0.05, and 0.01 levels, respectively, with standard errors reported in parentheses.

	Whether Settled (Logit)			**Payout-to-Claim Ratio (OLS)**		
	(1)	(2)	(3)	(4)	(5)	(6)
Avg. Similarity	11.212***			1.107***		
	(1.561)			(0.423)		
Length		0.0002***			0.00001**	
		(0.000)			(0.000)	
Negative Tone			-2.889			-3.809***
			(3.986)			(1.306)
Other Controls	Y	Y	Y	Y	Y	Y
Obs.	3,183	3,183	3,183	3,016	3,016	3,016
Pseudo R-Square/R-Square	0.146	0.136	0.114	0.012	0.011	0.013