Text mining letters from financial regulators to firms they supervise

David Bholat and James Brookes
Advanced Analytics, Bank of England, UK

Abstract

Our article uses text mining techniques to examine confidential letters sent from the Bank of England's Prudential Regulation Authority (PRA) to financial institutions it supervises. These letters are a 'report card' written to firms annually, and are the most important, regularly recurring written communication sent from the PRA to firms it supervises. Using two complementary machine learning techniques—random forests and logistic ridge regression—we explore whether the letters vary in substance and style depending on the size and importance of the firm to whom the PRA is writing. We find that letters to high impact firms use more evaluative, judgment-based language, and adopt a more forward-looking perspective. We also examine how PRA letters differ from similarly purposed letters written by its predecessor, the Financial Services Authority. We find evidence that PRA letters are different, with a greater degree of forward-looking language and directiveness, reflecting the shift in supervisory approach that has occurred in the UK following the financial crisis of 2007–09.

Correspondence:
James Brookes, Bank of England, Threadneedle Street, London EC2R 8AH, UK.
E-mail:
james.brookes@bankofengland.co.uk

1 Introduction

This article presents results from text mining confidential Periodic Summary Meeting (PSM) letters sent annually to UK banks and building societies by the Prudential Regulation Authority (PRA)—the microprudential[1] financial regulatory authority in the UK. These letters are the single most important, regularly recurring communication from UK financial regulators to the firms they supervise. In this article, we have constructed and measured several linguistic and discursive features of PSM letters to explore the extent to which supervisory communications vary depending on the size and significance of the firm to which they are written. This is an important empirical exercise to undertake given concerns, on the one hand, that some firms are considered 'too big to fail' and unfairly receive preferential treatment from regulators (Dowd *et al.*, 2011; Kane, 2016) and, on the other, the emphasis regulators place on fairness and proportionality (Bank of England, 2018). We also assess the extent to which PSM letters differ from the Advanced Risk-Responsive Operating frameWork (ARROW) letters written by the previous UK financial regulator, the Financial Services Authority (FSA), which was a separate organization from the Bank of England and was disbanded in the wake of the financial crisis. Examining letters written by the two regulators is also an important exercise to undertake in order to assess whether there truly has been a regime shift after the financial crisis in how UK financial firms are regulated, at least in terms of how regulators communicate with firms. This matters because, as we detail later, inadequate communication from financial regulators to firms they supervised was a contributory cause of the last financial crisis.

We see our article as contributing to the blossoming literature investigating central bank and financial regulatory communications. This literature has flourished recently for two reasons. First, during and after the financial crisis, communication

doi:10.1093/llc/fqz063 Advance Access published on 13 December 2019

emerged as a key regulatory tool for stabilizing the financial system once interest rates neared zero and could not practically be lowered much further.[2] For example, Mario Draghi's famous declaration that the European Central Bank would 'do whatever it takes' was arguably the single most important intervention that tempered the Eurozone crisis (Bholat *et al.*, 2018). As a result, much of the literature on central bank and financial regulatory communications focuses on how their public pronouncements impact financial markets (Blinder *et al.*, 2008; Brand *et al.*, 2010; Conrad and Lamla, 2010; Ranaldo and Rossi, 2010; Hayo *et al.*, 2012).

A second and related factor for the growing literature on central bank and financial regulatory communications has to do with public accountability. Because the financial crisis greatly increased their powers, central bankers and other financial regulators have increased the frequency of their communication and made efforts to improve its clarity to give greater public transparency to the rationale behind regulations. As the former Fed Chair Janet Yellen noted, central banks and financial regulators have over time moved from a position of 'never explaining' their policy rationale to a position where often 'the explanation *is* the policy' (Yellen, as quoted in Holmes, 2013). As a result, there is increasing interest among researchers in how central banks optimize their communication to ensure their continued democratic legitimacy with the general public (Hansen *et al.*, 2017; Bholat *et al.*, 2018; Haldane and McMahon, 2018).

Looking across the literature, most researchers have had to rely, by necessity, on public communications made by central banks and financial regulators because this is the only data that are available to them. Our contribution in this article is to examine private communications. In so doing, our article provides much needed empirical insight on the relationship between financial institutions and their regulators which, in the existing literature, is too often discussed based on theoretical priors, with some hypothesizing that the relationship between supervisors and the firms they regulate is inevitably too cozy (Kane, 2016), while others see the relationship as intrinsically antagonistic (Dowd and Hutchinson, 2014). As our article shows, neither of these theoretical priors is particularly well supported by the empirical evidence.

Besides being of interest to scholars of banking supervision and financial communication, we believe that our research methodology will interest academics empirically examining the relationships between regulators and the industries they regulate. For example, there are numerous other sectors besides finance, where poor regulatory communication has played a part in catastrophes, such as the 2010 Deepwater Horizon oil spill in the Gulf of Mexico (National Commission on the BP Deepwater Horizon Oil Spill and Offshore Drilling, 2011) and the Buncefield storage depot explosion in the UK in 2005 (Buncefield Major Incident Investigation Board, 2012). The methodology we develop in this article is modular and could thus be applied by scholars of other regulatory sectors in their own research.

The structure of our article is as follows. In Section 2, we discuss the PRA's overall approach to supervision, and the role that PSM letters play within it. Suffice it to say here that PSM letters are fundamental to how financial supervision is enacted in the UK. In Section 3, we discuss the data used in our research and then define and motivate the linguistic features we extract and measure from the letters. Section 4 details our methodology. We use two complementary techniques: (1) random forests, a machine learning algorithm in which classifications from many decision trees are averaged and (2) logistic ridge regression, a technique that can handle issues of multicollinearity[3] and perfect separation[4] among predictor variables that are usually insuperable within a standard logistic regression framework. Here, we contribute to the growing machine learning literature by providing an intuitive explanation of algorithms often described as 'black boxes'. Sections 5 and 6 then present our random forest and logistic ridge regression analyses of differences in communicative styles of the pre- and postcrisis regulators, and to different types of firms, respectively. Section 7 concludes. Looking at our findings in the round reveals that communications from UK financial regulators to the firms are not cut from a single cloth. This implies that 'structural' theories of financial regulation that depict the relationship in broad-brush terms miss important nuances in how the communications that constitute that relationship vary across time and across firms. Our contribution in this article is to identify the linguistic and discursive features that mark this variety.

2 Background

The global financial crisis first manifested itself in the UK with the run on Northern Rock. For most of its history, Northern Rock was a small, member-owned building society that operated in North East England. By the early 2000s, however, it had become the fifth largest mortgage lender in the UK using funds borrowed internationally to support its growth (Bholat and Gray, 2013). While not the main cause, inadequate bank supervision played a role in the firm's failure. In a postmortem audit, the FSA, the regulatory body overseeing Northern Rock at the time, specifically highlighted problems with its supervisory communications to the firm. For example, some messages were passed to the firm before they had been approved internally (Financial Services Authority Internal Audit Division, 2008). On other occasions, some messages that should have been communicated were not, for example, about the inadequacy of its stress testing process. Overall, the audit found that the FSA had not been sufficiently clear in its official communications with the firm. Similar issues with supervisory communications played themselves out in the lead up to crises at HBOS and the Royal Bank of Scotland (Financial Services Authority, 2011; Financial Conduct Authority and Prudential Regulation Authority, 2015). Partly because of these failings, the FSA was disbanded, and prudential banking supervision handed over to the PRA, part of the Bank of England. The PRA's statutory objectives are to promote the safety and soundness of the firms it regulates, and to contribute to the securing of an appropriate degree of protection for insurance policyholders, alongside a secondary objective to facilitate effective competition.

A key milestone in any given year for PRA supervisors is the PSM. These are annual meetings where supervisors agree the key risks posed by the firm to the PRA's objectives; look back at supervisory work conducted over the past 12 months; approve the proposed supervisory plan for the next 12 months; and reassess the longer term supervisory strategy for the firm. Each PSM meeting involves the supervisory team responsible for overseeing the firm, presenting to a PSM panel composed of senior PRA leaders not directly involved in the supervision of the firm. The PSM panel's role is to provide independent feedback and challenge on the proposed supervisory strategy and messages.[5]

The nature of the PSM depends on a firm's category. The PRA puts firms into five broad categories. Category 1 firms are those with a significant capacity to cause major disruption to the UK financial system. At the opposite end of the spectrum, Category 5 firms are those with almost no capacity to cause disruption to the UK financial system (Table 1).

Reflecting its 'proportionate' approach to regulation (Bank of England, 2018), the PRA supervises Category 1 firms more intensively than lower category firms. While a single Category 1 firm will be supervised by multiple supervisors, multiple Categories 4–5 firms will be supervised by a single supervisor. Thus, for Category 1 firms, the PSM is convened at the most senior level of the PRA. Outcomes of Category 1 PSM meetings are also shared with the Prudential Regulation Committee (PRC, formerly the PRA Board). For Categories 1, 2, and 3 firms, PSMs are firm-specific meetings. For Categories 4 and 5 firms, the PSM may consider groups of firms in the same category together, although each firm is still considered on a case-by-case basis.

The outcomes of the PSM meetings are communicated to firms via a PSM letter sent afterward.

Table 1 PRA firm categories

Category 1	Most significant deposit-takers capable of very significant disruption
Category 2	Significant deposit-takers capable of some disruption
Category 3	Deposit-takers capable of minor disruption
Category 4	Deposit-takers capable of very little disruption
Category 5	Deposit-takers capable of almost no disruption

Broadly speaking, the PSM letter is intended to convey the PRA's judgment of the most material risks facing firms. Ultimately, it is meant to drive change by the firm to mitigate these. The PSM letters are therefore drafted with care, and often redrafted and reviewed many times by different stakeholders to ensure that the messages prompt corrective actions. Although there is no style guide as such, there is a stockpile of existing PSM letters that are sometimes used as baselines for drafting new ones.[6]

While all PRA firms receive a PSM letter, the level of any additional supervisory communication (which may elaborate and update the supervisory messages in the PSM letter) will vary depending on the firm category and risk profile. For the highest risk firms, additional supervisory communication could potentially be of similar importance to the PSM letter.

In sum, the PSM letters are the most important, regularly recurring written communication from UK financial regulators to the firms they supervise. The fact that supervisors and PRA senior management expend significant effort to draft and redraft them makes them ideal sources for intimate insight into the usually cloistered and closely guarded communication that occurs between regulators and regulated entities. In what follows, we analyze these letters.

3 Data and Features

3.1 Data

The data we use to delve into the regulatory process are the PSM letters sent by the PRA, and the similarly purposed ARROW letters sent by the FSA. For our postcrisis data, we focused our analysis on a representative sample of comparable UK banks and building societies supervised by the PRA, with 3 years of PSM letters (2014–16). To make our analysis of firm category more tractable, we partitioned firms into two groups—Category 1 firms versus Categories 2–4 firms. This partitioning reflects the business reality, whereby Category 1 firms are supervised by a separate directorate from Categories 2–4 firms. We excluded Category 5 firms altogether given that these are almost all credit unions. In the UK, credit unions are wholly different types of institutions from banks and building societies. Many, if not most, credit unions are essentially not-for-profit social enterprises that operate within a particular locale. These institutions are small and offer a very simple product range. The 450 credit unions assets combined account for just 0.07% of the assets of the UK deposit taking sector (Proudman, 2018). As a consequence, the regulation that pertains to them is different from that which pertains to other deposit-takers, that is to say, banks and building societies. Consequently, we excluded them from our analysis.

For our precrisis sample, we trawled through records of FSA supervisory correspondences in the years before 2007 and were able to gather a convenience sample of FSA ARROW letters addressed to a number of comparable UK banks and building societies. In total, we harvested $n = 220$ letters. Of the total letters, 75% ($n = 165$) were PRA PSM letters and 25% ($n = 55$) were FSA ARROW letters. In the PRA sample, 7.3% ($n = 12$) of the letters were sent to Category 1 firms and 92.7% ($n = 153$) were sent to Categories 2–4 firms.[7]

3.2 Linguistic features

One approach to analyzing the letters would be using qualitative methods, such as close reading. However, this can be subjective and cannot be easily applied at scale. Thus, we adopt quantitative techniques from natural language processing (NLP; Jurafsky and Martin, 2014) and computational text analysis (Jockers, 2014), which can be applied more objectively and at scale.

For all the letters, we constructed a set of 11 linguistic and discourse features. Each of these features is intended to capture aspects of the PRA's stated approach to supervision (Bank of England, 2018).[8] For the most part, we employed a lexicon-based approach to feature engineering in which we used counts of specific words occurring in the letters normalized for the length of the letter. These lexicons were developed in three ways: (1) by drawing on previous literature relating to finance and supervision; (2) by gleaning words and phrases from the PRA's intranet pages and internal- and external-facing publications; and (3) by consulting external experts in the areas of linguistic and discourse analysis.[9] These features are summarized below.

3.2.1 *Forward-lookingness*

The precrisis model of supervision was criticized ex post for relying too heavily on backward-looking indicators of financial performance such as accounting ratios—an approach that proved inadequate for predicting emerging risks facing firms (Financial Services Authority, 2009; Viñals and Fiechter, 2010; Kellermann *et al.*, 2013). Postcrisis, many financial regulators around the world have adopted a 'forward-looking' approach, including De Nederlandsche Bank (DNB) and the UK's PRA. According to the PRA, being forward-looking means assessing 'firms not just against current risks, but also against those that could plausibly arise in the future' (Bank of England, 2018, p. 8). As a result, business model analysis and stress testing play an increasingly important role in how the PRA supervises firms (Breckenridge *et al.*, 2014; Dent *et al.*, 2016). We would thus expect to find more forward-lookingness in the postcrisis letters than in the precrisis letters.

Forward-looking sentences: Drawing on prior NLP studies that have examined the use of forward-looking language in corporate filings (Li, 2010; Bozanic *et al.*, 2018), we annotated a feature for the proportion of sentences in a letter that both contained a future-oriented term and had no past-tense form. Our list of future-oriented terms includes words such as: '*ahead*', '*aim*', '*anticipate*', '*approaching*', '*forthcoming*', '*future*', '*goal*', '*impending*', '*long-term*', '*outlook*', '*potential*', '*predict*', and '*project*'.

3.2.2 *Judgment*

As noted by De Nederlandsche Bank (2010), precrisis supervision was criticized for being a box-ticking activity. To put it in terms familiar to students of the central banking literature, the emphasis was on applying rules, while limiting the scope of supervisory discretion. Postcrisis, there has been a paradigm shift, with regulatory agencies stressing the importance of supervisory judgment (Andersson *et al.*, 2013; Lastra, 2013). For its part, the PRA explicitly states that judgment lies at the heart of its supervisory strategy (Bank of England, 2018). So we would expect to find more language around judgment in the PRA letters compared with the FSA letters. We engineered two features relating to this dimension of postcrisis prudential supervision:

- Evaluative words: our first measure examines the relative amount of evaluative language in a letter. We operationalized this by counting the number of words conveying an affective (positive or negative) stance in a financial context, for which we draw on the lexicons described and utilized in prior financial sentiment research (Loughran and McDonald, 2011; Bodnaruk *et al.*, 2015; Loughran and McDonald, 2016).[10] Affective words are words, such as: '*contentious*', '*(in)adequate*', '*inconvenient*', '*(in)effective*', '*spectacular*', '*stable*', and '*taint*'.
- Judgment verbs: second, we measured the rate at which explicit expressions of judgment such as '*we/PRA/FSA judge/consider/believe/deem*' occur in a letter.

3.2.3 *Directiveness*

Precrisis regulators were criticized for taking a 'light-touch' approach with financial firms (Financial Services Authority, 2009; Dowd *et al.*, 2011). This included not challenging senior management (Viñals and Fiechter, 2010; Sijbrand and Rijsbergen, 2013); not intervening firmly enough to ensure issues were addressed (Garcia, 2009); and not probing deep enough to identify latent risks (Kellermann and Mosch, 2013). Thus, Sijbrand and Rijsbergen (2013, p. 22) argue that postcrisis supervision 'needs to be more challenging, intrusive and comprehensive in its approach'. The following four indicators are designed to detect whether supervision has become more directive:

(1) Directive words: our first feature here is the number of directives in a letter. We measure this by the rate of word stems with obligative meaning, e.g. '*expect*', '*need*', '*should*', '*require*', '*request*', '*ask*', '*must*', '*want*', and '*ought*'.

(2) Letter length: we used the length of the letter to proxy the amount of detail the supervisor

goes into, which may indicate a more probing analytical approach.

(3) Negative words: we measured the rate of negative words in a letter (drawn from the financial sentiment lexicon mentioned earlier).

(4) Deadlines: precrisis supervision was criticized for leaving time frames up to firms' senior management (Kellermann and Mosch, 2013). So we counted the number of deadlines in a letter. If a request is given a deadline, it is more pointed than one that is not. If there is no deadline, a firm may feel less urgency to oblige with the supervisory request because it is seen as being less urgent. From each letter, we extracted all substrings that matched '*by*' followed by a date formulation (e.g. '*31st January*', '*the end of January*', '*end January*', '*end Q1*', etc.) and counted them.

3.2.4 *Distance*

After the financial crisis, regulators were blamed in some quarters for being too close to the entities they supervised, especially systemically important firms. The most strident voices have even spoken of 'regulatory capture' (Baker, 2010; Capuder, 2011; Dowd *et al.*, 2011; Johnson and Kwak, 2011). Postcrisis, supervisory bodies such as the DNB have explicitly adopted a rotation policy, whereby supervisors spend no more than 4 or 5 years supervising a firm (Kellermann and Mosch, 2013). Rotation is aimed at ensuring that supervisors maintain 'an arm's-length relationship' (Viñals and Fiechter, 2010, p. 17), thereby reducing the potential for regulatory capture. Consequently, we expected to find more linguistic formality in the PRA letters, which would indicate a more 'arm's-length' relationship between the PRA and regulated firms.

We derived two features indicative of regulatory distance:

(1) Salutation style: first, an explicit indicator of the degree of formality in a letter is the style of salutation. A letter whose salutation is addressed to an individual with their first name is much less formal than one addressed to a generic collective, for instance, a firm's board. Furthermore, a letter that has a handwritten salutation is less formal than one that is typed. Thus, a letter whose salutation is 'Dear Jennifer' is more informal, friendly, and suggestive of a closer relationship than the more aloof and distant '`Dear Board Members`'.

(2) Personal pronouns: second, we measure the local personal pronoun rate—first person pronouns ('*I*', '*me*', '*my*', '*mine*', '*we*', '*us*', '*our*', and '*ours*') and second person pronouns ('*you*', '*your*', and '*yours*'). We take an abundance of personal pronouns to indicate a close-knit relationship between regulator and firm (for theoretical motivation see Biber, 1991).

3.2.5 *Systems-thinking*

Another criticism leveled against supervisors in the wake of the financial crisis is that they had supervised institutions in isolation, without considering sector-wide themes or institutions' interconnectedness in any great detail (Sijbrand and Rijsbergen, 2013). In his review of the factors contributing to the global financial crisis, Adair Turner, the former chair of the FSA, noted it to be 'a common feature, and in retrospect a common failing, of bank regulation and supervisory systems across the world' (Financial Services Authority, 2009, p. 87). To prevent this happening again, Turner recommended '[a]n increase in resources devoted to sectoral and firm comparator analysis, enabling the FSA to better identify firms which are outliers in terms of risks and business strategies and to identify emerging sector-wide trends which may create systemic risk' (Financial Services Authority, 2009, p. 88). As a way of influencing the firm to reduce risk-taking behavior, this may constitute an effective rhetorical technique. According to Richards (2013: p. 89), '[o]ne of the most effective means of persuasive supervision is for supervisors to provide the institution's board and management with observations as to the institution's risk profile relative to peers. Supervisors can benchmark key risk profile indicators and provide feedback to the bank on how it compares to peers.'

To assess the extent to which supervisors adopt a macroprudential or systemic point of view, we configured two indicators of this style:

(1) Systems-thinking words: first, we counted words indicating that a comparison was being made to other entities, e.g. '*outlier*', '*compare*', '*sector*', '*benchmark*', '*peers*', '*trends*', '*sector-wide*', '*systemic*', and '*global*'.
(2) Comparative modifiers: second, we counted all tokens part-of-speech tagged[11] as a superlative or comparative adjective or adverb, such as '*healthiest of firms*', '*more X than Y*'.

Table 2 presents a summary of the quantitative features across our dimensions of interest. Two initial observations are of note. First, linguistic features indicative of forward-lookingness and judgment are, on the whole, more common in PRA letters compared with FSA letters. PRA letters are also longer, more stern in sentiment and show greater evidence of systems-thinking. Second, among PRA letters, there is relatively more forward-lookingness and evaluation—but fewer deadlines and directives—in letters to Category 1 firms compared with letters to Categories 2–4 firms. In the following sections, we go beyond these casual descriptive statistics to more rigorously modeling the relationship between linguistic features, and firm category and regulatory agency.

4 Methodology

To understand the relationship between the linguistic features and to whom (firm category) and when (before the financial crisis by the FSA versus after the crisis by the PRA) they were sent, we used a machine learning algorithm known as random forests. We then triangulated our findings from the random forest analysis using a regularized[12] version of logistic regression called logistic ridge regression.[13] We explain these two methods below.

4.1 Random forests

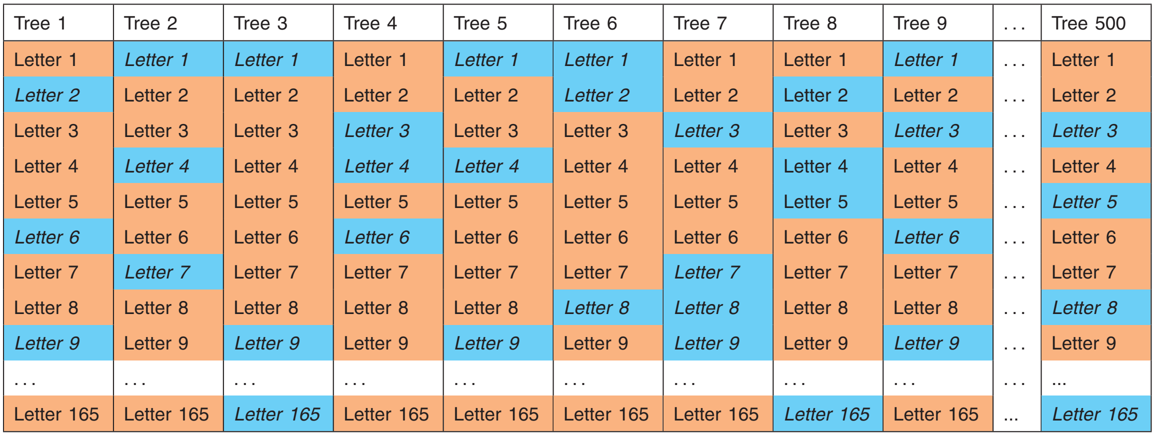
Random forests have been applied in a variety of disciplines, e.g. in ecology, to model the presence of invasive plant species (Cutler *et al.*, 2007); in linguistics, to model individuals' grammatical choices (Tagliamonte and Baayen, 2012; Baayen *et al.*, 2013); in bibliometry, for authorship disambiguation (Treeratpituk and Giles, 2009); and in energy studies, to better understand household fuel choices (Vahlne, 2017).

Our random forest model worked roughly as follows. We started from the full set of letters and their features, with the random forest algorithm then drawing random subsamples of letters 500 times. Each time, i.e. for each decision tree, about 63% of the letters were included in a randomly sampled training set. The remaining 37% of observations not

Table 2 Summary of quantitative variables

Dimension	Feature	PRA				FSA	
		Category 1		Categories 2–4			
Forward-lookingness	Forward-looking sentences (%)	35	(6)	32	(9)	26	(7)
Judgment	Evaluative words	36,653	(8,019)	25,959	(6,356)	18,252	(5,881)
	Judgment verbs	1,547	(931)	940	(1,001)	1,016	(987)
Directiveness	Directives	13,724	(2,391)	19,347	(4,689)	14,142	(3,298)
	Letter length	1,890	(470)	1,488	(501)	1,228	(592)
	Negative words	22,316	(8,040)	15,650	(4,936)	11,935	(4,126)
	Deadlines	506	(343)	1,871	(1,318)	1,132	(650)
Distance	Personal pronouns	33,960	(13,558)	35,636	(11,174)	49,020	(16,068)
Systems-thinking	Systems-thinking words	649	(754)	715	(995)	442	(837)
	Comparative modifiers	5,643	(2,199)	3,891	(2,193)	3,057	(1,875)

Notes: Mean and, in parentheses, standard deviation. All features, except for letter length and forward-looking sentences, are normalized for letter length and multiplied by 10^6 to give a count per million, a common normalization method in the corpus linguistics literature (McEnery and Hardie, 2012).



Tree 1	Tree 2	Tree 3	Tree 4	Tree 5	Tree 6	Tree 7	Tree 8	Tree 9	. . .	Tree 500
Letter 1	*Letter 1*	*Letter 1*	Letter 1	*Letter 1*	*Letter 1*	Letter 1	Letter 1	*Letter 1*	. . .	Letter 1
Letter 2	Letter 2	Letter 2	Letter 2	Letter 2	*Letter 2*	Letter 2	Letter 2	Letter 2	. . .	Letter 2
Letter 3	Letter 3	Letter 3	*Letter 3*	Letter 3	Letter 3	*Letter 3*	Letter 3	*Letter 3*	. . .	*Letter 3*
Letter 4	*Letter 4*	Letter 4	*Letter 4*	*Letter 4*	Letter 4	Letter 4	Letter 4	Letter 4	. . .	Letter 4
Letter 5	Letter 5	Letter 5	Letter 5	Letter 5	Letter 5	Letter 5	Letter 5	Letter 5	. . .	*Letter 5*
Letter 6	Letter 6	Letter 6	*Letter 6*	Letter 6	Letter 6	Letter 6	Letter 6	*Letter 6*	. . .	Letter 6
Letter 7	*Letter 7*	Letter 7	Letter 7	Letter 7	Letter 7	*Letter 7*	Letter 7	Letter 7	. . .	Letter 7
Letter 8	Letter 8	Letter 8	Letter 8	Letter 8	*Letter 8*	*Letter 8*	Letter 8	Letter 8	. . .	*Letter 8*
Letter 9	Letter 9	*Letter 9*	Letter 9	*Letter 9*	Letter 9	*Letter 9*	Letter 9	*Letter 9*	. . .	Letter 9
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	...
Letter 165	Letter 165	*Letter 165*	Letter 165	Letter 165	Letter 165	Letter 165	*Letter 165*	Letter 165	...	*Letter 165*

Fig. 1 Example of training and test data. Training set observations are given in normal font and test set observations are given in italics. The first column means that ∼63% of the letters sampled as the training set for the first tree include letters 1, 3, 4, 5, 7, 8, and 165 (in normal font); the 37% of the letters that are left out as the test set include letters 2, 6, and 9 (in italics). Similarly, the second column means that letters 2, 3, 5, 6, 8, 9, and 165 are included in Tree 2's training set (in normal font), while letters 1, 4, and 7 are included in Tree 2's test set (in italics)

included made up the test set for the tree.[14] Fig. 1 illustrates the sampling procedure for a portion of the category data. Note that here, we show training set observations in normal font and test set observations in italics.

A decision tree was then built for each of the 500 training sets, splitting first on the most important feature identified by the random forest algorithm in separating the letters—i.e. between Category 1 and Categories 2–4 letters in the case of firm category classification; and between pre-crisis FSA letters and postcrisis PRA letters in the case of regulator classification. The decision tree algorithm then continually subdivided the data along linguistic features of successively diminishing discriminative power until there was a single observation in the terminal node (branch) or the data could not meaningfully be subdivided by features any further. The number of features considered at each split point was the square root of the number of features, called *mtry*. In our models, three randomly sampled features were considered at each split point, i.e. the rounded down square root of our full set of 11 features. This procedure induced more randomness, making the trees more diverse. In doing so, this ensured that our model did not overfit to the training data and could generalize out of sample. The randomness of the letters selected for any given tree and the randomness of the features considered at any given split point are why the collection of these trees is termed a random forest.

To make the modeling approach clear, consider the following example. Fig. 2 shows (part of) one of our decision trees.[15] The decision tree algorithm randomly sampled three features from the full set of 11 at its first branch (the ones tagged as Node 1). The algorithm then identified which of these features significantly distinguished letters written to Category 1 firms from those written to Categories 2–4 firms. In this tree, the type of salutation (informal versus formal) was the most important feature. If this were a regression, we might say that this is the feature with the largest coefficient. Hence, this feature appears at the highest point in the tree's hierarchy. At split ③, another random selection of three features was sampled from the 11 features (the ones tagged as Node 3). The algorithm identified that for letters with formal salutations, letter length is an important determinant, with longer letters (>1,500 words) being sent to mostly Category 1 firms. This splitting process continued until the criterion mentioned above was reached (data not shown).

After learning to identify the discriminating linguistic features from the training data, the tree was

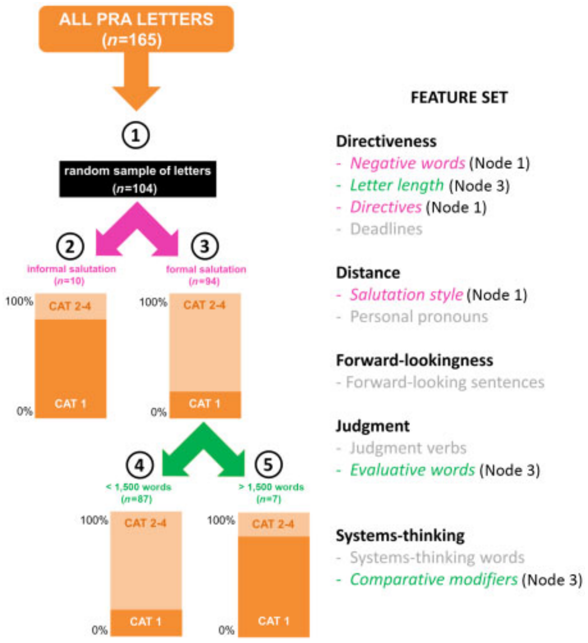


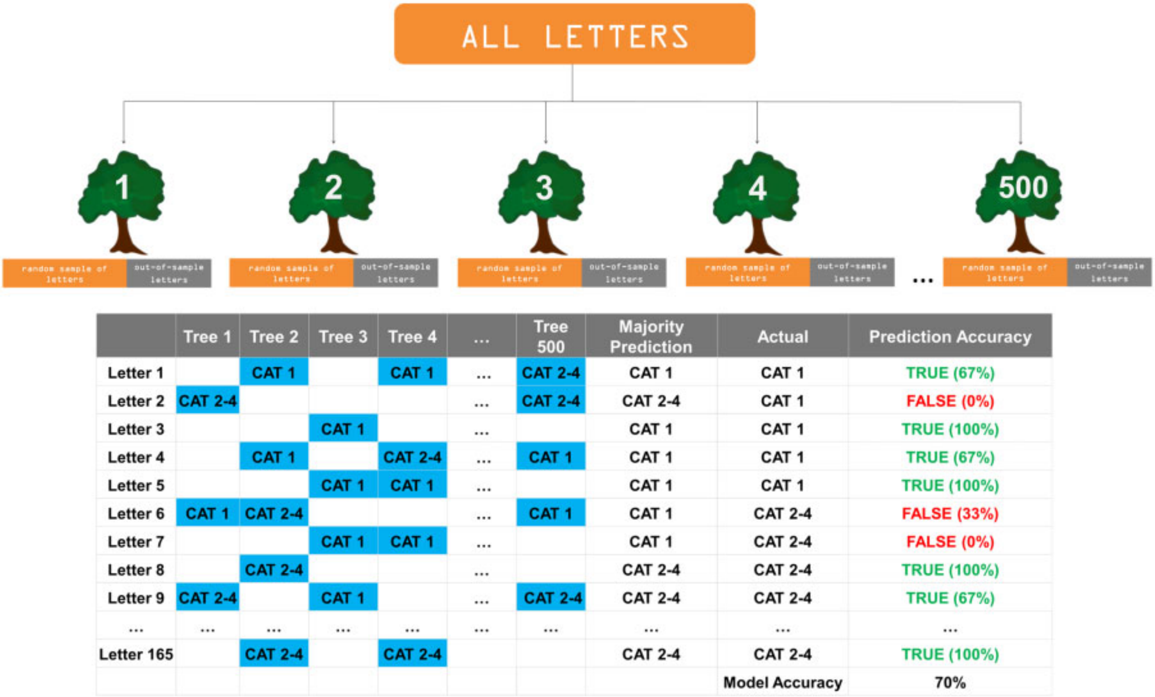
Fig. 2 (Part of) one of the trees in our random forest using *mtry* = 3. At the first node, the italicized features tagged as Node 1 were randomly sampled. At node 3, the italicized features tagged as Node 3 were randomly sampled. In this tree, the grayed-out features in normal font were not sampled at all

used to classify the remaining letters in the test set. This was how we assessed the model's accuracy. For each of the remaining letters, the algorithm made a prediction; in the above example, whether the letter is to a Category 1 firm or not. For example, if one of these out-of-sample letters had a formal salutation and a letter length greater than 1,500 words, it would be predicted to be a Category 1 firm letter.

We grew 500 trees this way. The result is a 'random forest' of decision trees. For any given letter, the predictions for when that observation was out-of-sample were combined to produce a majority prediction for that letter. Then, the accuracy of the random forest is the proportion of times the majority prediction matched reality (Fig. 3). Thus, looking at Fig. 3, Letter 1 is out-of-sample in Tree 2 (prediction: Category 1 firm), Tree 4 (prediction: Category 1 firm), Tree 500 (prediction: Categories 2–4 firm). Here, two-thirds of the predictions for Letter 1 are Category 1, so that class is the forest's mode prediction for Letter 1. Note that the prediction for Letter 1 matches its actual class. It is an accurate prediction. Then, the accuracy of the random forest as a whole is the proportion of times the majority prediction matches reality; for this example random forest, the accuracy is $\frac{7\text{ matches}}{10\text{ letters}} = 70\%$ because 7 out of 10 letters were correctly identified.

We assessed the relative influence of our 11 features in this model with a variable importance measure. Simplifying somewhat, each feature was given a score based on how much, on average, the predictive accuracy of a single tree in the forest dropped when the feature was shuffled. If out-of-sample accuracy dropped a lot, then the feature was deemed useful at discriminating between letters. If the drop in accuracy was negative or close to nil, this suggested that the feature's importance was low. Appendix A provides a worked through example.

The inherent randomness involved in both sampling the letters and sampling the features for splitting means we are likely to get different results between any two random forests runs. Thus, to ensure that our random forest model and the



	Tree 1	Tree 2	Tree 3	Tree 4	...	Tree 500	Majority Prediction	Actual	Prediction Accuracy
Letter 1		CAT 1		CAT 1	...	CAT 2-4	CAT 1	CAT 1	TRUE (67%)
Letter 2	CAT 2-4				...	CAT 2-4	CAT 2-4	CAT 1	FALSE (0%)
Letter 3			CAT 1		...		CAT 1	CAT 1	TRUE (100%)
Letter 4		CAT 1		CAT 2-4	...	CAT 1	CAT 1	CAT 1	TRUE (67%)
Letter 5			CAT 1	CAT 1	...		CAT 1	CAT 1	TRUE (100%)
Letter 6	CAT 1	CAT 2-4			...	CAT 1	CAT 1	CAT 2-4	FALSE (33%)
Letter 7			CAT 1	CAT 1	...		CAT 1	CAT 2-4	FALSE (0%)
Letter 8		CAT 2-4			...		CAT 2-4	CAT 2-4	TRUE (100%)
Letter 9	CAT 2-4		CAT 1		...	CAT 2-4	CAT 2-4	CAT 2-4	TRUE (67%)
...	...	...	...	...	...	...	...	...	...
Letter 165		CAT 2-4		CAT 2-4			CAT 2-4	CAT 2-4	TRUE (100%)
								Model Accuracy	70%

Fig. 3 Random forest model and predictions

resulting feature importance rankings were robust, we built 1,000 forests and averaged results.[16]

4.2 Logistic ridge regression

To triangulate our results from the random forest, we used a logistic ridge regression. Specifically, we used a pair of binary logistic regression models, where the target or outcome variables we wanted to predict were regulator (FSA versus PRA, as a proxy of pre- and postcrisis regulation) and firm category (Category 1 versus Categories 2–4). In formal terms, respectively,

$$\log\left(\frac{P(\text{PRA letter})}{1-P(\text{PRA letter})}\right) = \beta_0 + \beta_1\text{LetterLength} + \beta_2\text{Deadlines} + \ldots + \beta_{11}x_{11}, \qquad (1)$$

$$\log\left(\frac{P(\text{Category 1 letter})}{1-P(\text{Category 1 letter})}\right) = \beta_0 + \beta_1\text{LetterLength} + \beta_2\text{Deadlines} + \ldots + \beta_{11}x_{11} \qquad (2)$$

where $P(Y)$ is the probability of Y occurring, β_0 is the y-intercept, and the other β (betas) are the coefficients corresponding to our linguistic features.

In a standard logistic regression, the coefficients are estimated using maximum likelihood, $\ell^{\text{MLE}}(\beta)$. In short, this means searching for the set of estimates, which result in predicted probabilities that are as close as possible to the observed classes. Thus, for two response classes dummy coded as 1 and 0, this means obtaining predicted probabilities that are close to 1 and 0, respectively. More formally,

$$\ell^{\text{MLE}}(\beta) = argmax\left\{\sum_{i=1}^{n}\left[y_i \log\left(\frac{\widehat{P(y_i)}}{1-\widehat{P(y_i)}}\right) + \log\left(1-\widehat{P(y_i)}\right)\right]\right\}. \qquad (3)$$

Unfortunately, maximum likelihood estimation is problematic with data such as ours. The first issue is multicollinearity. For instance, in the category dataset, 8 features have a variance inflation factor

(VIF) well over 10, meaning they have a strong linear relationship with each other. As in a standard linear regression, such strong correlation among the features makes it difficult to ascribe variation in the target variable separately to each of the features and so their standard errors tend to be large. Second, and related, some features are on their own perfect predictors. For example, only informal salutations are written to Category 1 firms in our dataset. Such perfect separation also creates large standard errors.

To address both these issues, we include a ridge regression penalty term $\lambda \sum_{j=1}^{p} \beta_j^2$ within the loss function to estimate the model coefficients for the predictor variables:

$$\ell^{\mathrm{Ridge}}(\beta) = \mathop{argmax}\left\{ \sum_{i=1}^{n} \left[y_i \log\left(\frac{\widehat{P(y_i)}}{1-\widehat{P(y_i)}} \right) + \log\left(1-\widehat{P(y_i)}\right) \right] - \lambda \sum_{j=1}^{p} \beta_j^2 \right\}. \tag{4}$$

The λ (lambda) is a tuning parameter that controls the impact of the penalty. It is greater than or equal to zero. If $\lambda = 0$, the penalty term disappears so that $\ell^{\mathrm{Ridge}}(\beta) = \ell^{\mathrm{MLE}}(\beta)$; as λ grows, we get smaller coefficients that get closer and closer to zero. This penalizes coefficients that are far from zero and shrinks correlated variables with large coefficients (in absolute value). We established the optimal value for λ for our datasets using *k*-fold cross-validation over a grid of values for λ (we pass over the technicalities here, instead see James *et al.* (2013) for exquisite discussion). Note that, as is customary in ridge regression modeling, we standardized all predictors by subtracting the mean and dividing by the standard deviation, so that each coefficient is equally penalized and so that their relative strengths can be more easily compared (again, see James *et al.* (2013) for discussion).

By including a ridge regression penalty, this also reduces the likelihood of our model overfitting to the training data by shrinking the estimated β-coefficients of the linguistic features by the factor λ. While this means our model fits the training data less well than if there were no ridge regression penalty, it also means that our model likely better generalizes out-of-sample.

5 Postcrisis Changes in Supervisory Correspondence

One way to assess whether there has been a shift in supervisory approach after the financial crisis is to compare the PRA's PSM letters with a sample of FSA's ARROW letters written prior to 2007. We do so in this section.

Our random forest model predicting the regulator (FSA versus PRA) from the 11 linguistic features has a good fit, as gauged by the out-of-sample estimate. The model's accuracy is 89.52%, a considerable increase over the baseline of 75%, consistently predicting the overall most frequent occurring letter type in our data (here, PRA letters). However, because the data exhibit some class imbalance (25% FSA versus 75% PRA), we prefer to use another measure of performance. Specifically, we use Harrell's *C*-statistic of concordance (Harrell *et al.*, 1982)—equivalently the area under the receiver operating characteristic curve (AUC)—as a measure of how well the model discriminates between response classes.[17] The *C*-statistic is the probability, given the model, that a randomly chosen PRA letter will be assigned a higher predicted probability of being a PRA letter compared with a randomly chosen FSA letter. We clarify how this statistic is computed in Appendix B. The *C*-statistic of our random forest on the regulator data is 0.9374, meaning that the probability, given the model, that a randomly chosen PRA letter will be assigned a higher predicted probability of being a PRA letter compared with a randomly chosen FSA letter is well over 90%. In the statistical learning literature, a *C*-statistic as high as ours is considered 'outstanding' (for discussion, see, e.g. Hosmer *et al.*, 2013, p. 177).[18]

To identify influential variables in the random forest model of regulator, we used the variable importance measure as described earlier. In terms of interpretation, Strobl *et al.* (2009) suggest that for a single forest iteration, variables with importance scores that are negative, zero, or barely positive are uninformative and can be safely ignored. Formally, we state this threshold as $\theta = |\min(0, V)|$, where θ is the threshold below which a predictor is uninformative, and V is the set of all 11 importance scores for a given forest iteration. If the variable's value is above this threshold, the variable is considered informative. One can think of values above this threshold as being 'statistically significant' in an extremely loose sense of that phrase. However, given that we grew multiple forests, in order to focus solely on the best performing features across our 1,000 forest iterations, we modified this heuristic. Specifically, we considered a feature to be important if its minimum variable importance score over all the forest iterations exceeded the absolute value of the lowest variable importance score over all features and all iterations. Fig. 4 plots the results.

Fig. 4 shows that evaluative words, directives, and personal pronouns are the most important discriminators. For instance, permuting the feature for evaluative words results in an ~6% drop in accuracy. Other important features according to our measure are negative words, letter length, deadlines, forward-looking sentences, and comparative modifiers.

We subsequently fit a logistic ridge regression to triangulate the results of the random forests model. Table 3 gives the standardized coefficients associated with each feature, lower and upper bounds of their 95% confidence intervals, and exponentiated standardized coefficients.[19] Coefficients whose confidence intervals do not cross zero are considered relevant, and the features that they refer to are given in bold.

The logistic ridge regression results largely square with those from the random forests analysis, with some quirks (for instance, salutation style appears relevant in the logit model; letter length and negative words are absent). In what follows, we only

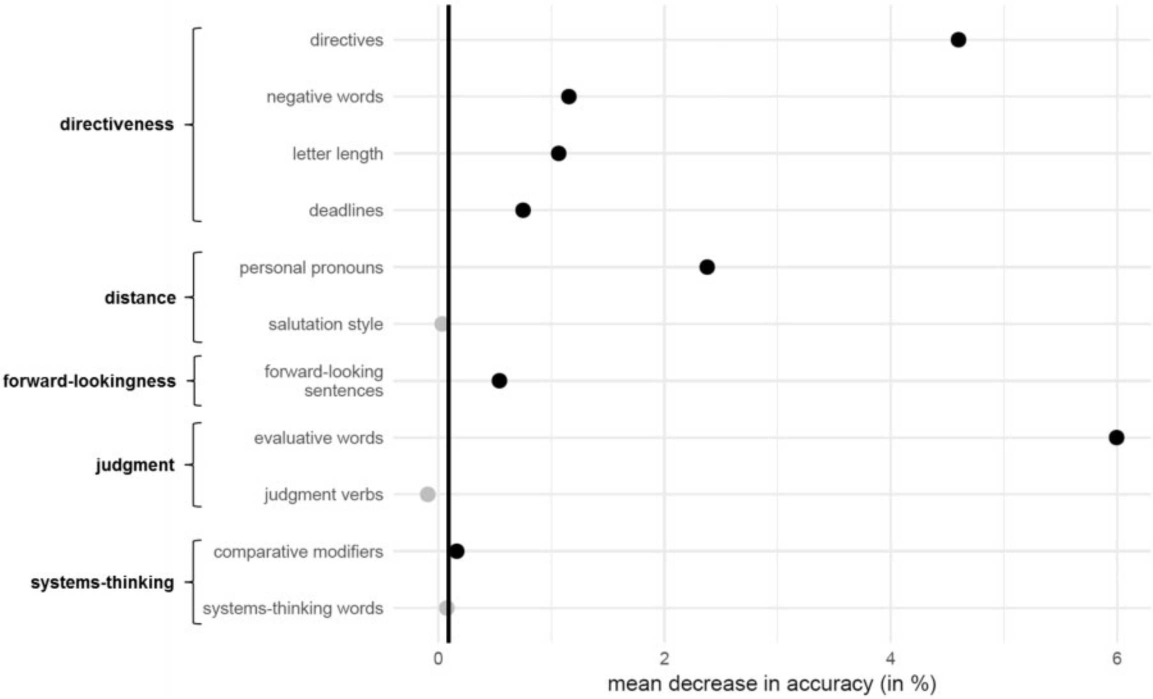


Fig. 4 Variable importance for the Regulator random forest model. The vertical line is the absolute value of the lowest variable importance score over all features and all 1,000 forest iterations. Features with scores to the right of the vertical line and colored black are considered influential. Features with scores on or to the left of the vertical line and colored gray are considered non-influential. Within each dimension of regulatory style, variables are sorted according their importance

Table 3 Logistic ridge regression results

			95% Confidence Interval		
Dimension	**Feature**	**Estimate**	**Lower**	**Upper**	**exp(Estimate)**
Directiveness	**Directives**	1.139	0.823	1.454	3.123
	Deadlines	0.656	0.326	0.986	1.927
	Negative words	0.276	−0.089	0.642	1.318
	Letter length	0.059	−0.311	0.429	1.061
Distance	**Salutation style**	0.303	0.137	0.469	1.354
	Personal pronouns	−0.617	−0.916	−0.318	0.540
Forward-lookingness	**Forward-looking sentences**	0.395	0.111	0.679	1.484
Judgment	**Evaluative words**	0.953	0.636	1.270	2.593
	Judgment verbs	−0.015	−0.286	0.257	0.985
Systems-thinking	Comparative modifiers	0.288	−0.031	0.608	1.334
	Systems-thinking words	−0.005	−0.306	0.295	0.995

Notes: We have modeled the log-odds of a letter being a PRA letter. Hence, positive coefficients are linked with PRA letters and negative coefficients with FSA letters. The salutation style predictor is a dummy variable, 0 if formal, 1 if informal. Variables were standardized prior to analysis. Thus, a point estimate for a feature indicates the increase/decrease in the log-odds for a one unit increase in the standard deviation of the predictor. The final column gives the exponentiated standardized coefficients (odds ratio). Features whose confidence intervals do not cross zero are given in bold.

discuss those features where both models agree on their importance.

5.1 Directiveness

The signs of the coefficients for directive expressions and deadlines are both positive, indicating that there are relatively more of them in the PRA letters as compared with the FSA letters. Looking at the odds ratios for these features—given in the rightmost column of Table 3—we see that a one standard deviation increase in directives (4,916 words per million) results in a 3.1-fold increase in the odds of being a PRA letter and, similarly, a one standard deviation increase in deadlines (1,220 words per million) results in a 1.9-fold increase in the odds of being a PRA letter.

5.2 Distance

The sign of our one important regulatory distance variable—the personal pronoun (‘*I*’, ‘*we*’, ‘*you*’) rate—is negative. The odds ratio for this feature suggests a 46% decrease in the odds of a letter being a PRA letter for a one standard deviation increase in the feature (13,930 words per million). We take this as evidence that, in general, the postcrisis regulator is keeping a greater critical distance between itself and regulated entities by adopting a more formal stance.

5.3 Forward-lookingness

The sign of the coefficient for forward-looking sentences is positive, meaning that as compared with the FSA, the PRA displays significantly more forward-looking judgment in its communication to firms. A one standard deviation increase in this feature (9 sentences per 100) results in a 48% increase in the odds of being a PRA letter. One of the critiques of the FSA was that it responded to risks only after they had crystallized. By contrast, the PRA aspires to be more proactive in its approach. Our analysis provides some evidence that PRA supervisors are indeed adopting such a style.

5.4 Judgment

The sign for evaluative words shows that as the amount of evaluation in a letter increases, it is more likely that such an example would be classified as a PRA letter. The effect of this feature is also sizable: for a one standard increase in the evaluative word rate (7,692 counts per million), we see a 2.6-fold increase in the odds of being a PRA letter. As this feature is a proxy for the amount of supervisory judgment, it would appear that PRA supervisors are exercising more judgment in correspondence with firms than did FSA supervisors.

6 Differences Between Different Category Firms

In this section we analyze whether the PRA communicates to firms the same irrespective of their size and significance. We would expect to find some stylistic differences here, given that one of the key tenets of the PRA's approach is to be a proportionate regulator (Bank of England, 2018). Again, we used the random forest machine learning algorithm to determine how distinctive the letter types are and to identify the important discriminating linguistic features, and then used logistic ridge regression to again triangulate our results.

Our random forest model predicting firm category (Category 1 versus Categories 2–4) from the 11 linguistic features has a very good performance, as gauged by the out-of-sample estimate. The model's accuracy is 97.28% (compared with the baseline of 92.73%, consistently predicting the most frequent class Categories 2–4) and the model's *C*-statistic of concordance, described above, is 0.9721 (recall that a score of 0.5 indicates a classifier that performs no better than chance).[20] In short, our random forest shows that Category 1 firm letters are on the whole stylistically different from those written to Categories 2–4 firms.

Our feature importance algorithm detects six features of relevance (Figure 5). Salutation style is by far the most impactful predictor, followed by deadlines, negative words, and evaluative words, then directives and comparative modifiers. Personal

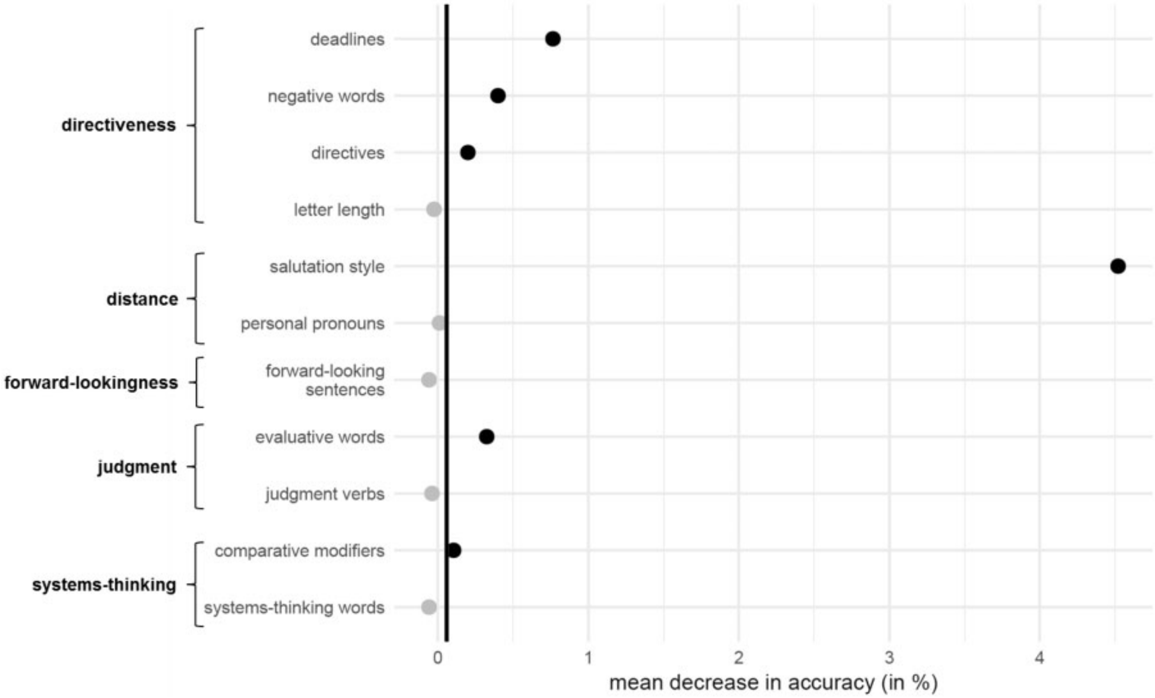


Fig. 5 Variable importance for the category random forest model. The vertical line is the absolute value of the lowest variable importance score over all features and all 1,000 forest iterations. Features with scores to the right of the vertical line and colored black are considered influential. Features with scores on or to the left of the vertical line and colored gray are considered non-influential. Within each dimension of regulatory style, variables are sorted according their importance

Table 4 Logistic ridge regression results

			95% Confidence Interval		
Dimension	**Feature**	**Estimate**	**Lower**	**Upper**	**exp(Estimate)**
Directiveness	Negative words	0.247	−0.150	0.644	1.280
	Letter length	0.087	−0.396	0.570	1.091
	Directives	−0.304	−0.471	−0.137	0.738
	Deadlines	−0.447	−0.646	−0.248	0.640
Distance	**Salutation style**	0.972	0.649	1.296	2.644
	Personal pronouns	−0.243	−0.604	0.119	0.784
Forward-lookingness	**Forward-looking Sentences**	0.218	0.029	0.407	1.244
Judgment	**Judgment verbs**	0.501	0.140	0.862	1.651
	Evaluative words	0.337	0.137	0.537	1.401
Systems-thinking	**Comparative modifiers**	0.401	0.092	0.711	1.494
	Systems-thinking words	−0.228	−0.466	0.010	0.796

Notes: We have modeled the log-odds of a letter being a Category 1 firm letter. Hence positive coefficients are linked with Category 1 firm letters and negative coefficients with Categories 2–4 firm letters. The salutation style predictor is a dummy variable, 0 if formal, 1 if informal. Variables were standardized prior to analysis. Thus, a point estimate for a feature indicates the increase/decrease in the log-odds for a one unit increase in the standard deviation of the predictor. The final column gives the exponentiated standardized coefficients (odds ratio). Features whose confidence intervals do not cross zero are given in bold.

pronouns, letter length, judgment verbs, systems-thinking words, and forward-looking sentences contribute little, if anything, to prediction.

We subsequently used a logistic ridge regression to triangulate the results of the random forests model. This model has a cross-validated accuracy of 98.79% and *C*-statistic of 0.9820, thus favorably comparable with the random forest model. Table 4 gives the standardized coefficients associated with each feature, lower and upper bounds of their 95% confidence intervals, and exponentiated coefficients. Coefficients whose confidence intervals do not cross zero are considered relevant, and the features that they refer to are given in bold.

Seven features are deemed relevant in the logistic regression model, compared with six in the random forest variable importance. Together, they agree on five important features: directives, deadlines, salutation style, evaluative words, and comparative modifiers. The importance in the logistic ridge regression results of forward-looking sentences and evaluative words, a proxy for judgment, is particularly noteworthy for two reasons. First, these features were also important discriminators of letters written by the PRA in general compared with the letters previously written by the FSA. Second, this provides some evidence that PRA supervision in practice is living up to its aspiration to be forward-looking and judgment-based.

7 Conclusion

In this article, we have quantitatively analyzed the text of PSM letters to understand how supervisory communications to firms has changed over time, and how it differs depending on the size and systemic importance of the firm with whom the PRA is communicating.

In summary, we have found a change in the communicative style of supervisors over time, reflecting the change in the authority responsible for micro-prudential regulation in the UK since the financial crisis. Our random forest suggests letters written by the PRA are distinct from those written previously by the FSA along a range of linguistic features. For example, our ridge regression analysis shows that PRA letters are relatively more forward-looking in their language. We also find that the PRA's letters are relatively more directive, formal, and judgment-based.

Second, we find differences in how the PRA communicates with firms of different sizes and significance, indicative of the PRA's proportionate approach to banking supervision. For example, the Category 1 firm letters are more evaluative and comparative. On the one hand, this may reflect that greater judgment is exercised toward firms which pose the greatest risk to financial stability, as well as the greater urgency to communicate to them macro-prudential issues, as one might expect from an efficient regulator. On the other hand, we find that the style of letters sent to systemic firms is generally less directive and less distant. Our best guess as to why this is the case is that the PRA has more frequent contact with the largest firms precisely because of their systemic importance. As a consequence, the PRA is more familiar with the senior management at those firms, and thus a less formal written style is appropriate.

Third, there are some commonalities between the models for regulator and the models for category. Specifically, three features consistently show up as being influential features in all four models—evaluative words, deadlines, and directive expressions. We believe the importance of these features is an empirical manifestation of the PRA's approach to banking supervision, namely, it aims to be a judgment-based regulator that expects action from the firm concerning the most material risks facing it. It is also worth noting that forward-lookingness shows up as being important in three of the four models (except the category random forests). Again, we believe this speaks to the PRA's emphasis on forward-looking analyses in its assessment of firms' risks.

Methodologically, we have shown how a text mining approach using machine learning can yield important insights on the day-to-day practices of supervision. Much research on supervision is based on case studies of the supervision of particular firms, such as the inquiries into Northern Rock (Financial Services Authority Internal Audit Division, 2008), the Royal Bank of Scotland (Financial Services Authority, 2011), and HBOS (Financial Conduct Authority and Prudential Regulation Authority, 2015). However, in some ways, case studies of individual institutions may not be representative of how most supervision is done. They may only yield insight on the instances where supervision had bad outcomes. Our approach is different. It can be considered a meta-analysis that generalizes over the supervisory styles toward individual institutions.

In closing, we offer a few thoughts about how the research we have conducted could be extended by other researchers. First, although our focus was on banks and building societies, the same PSM process applies to UK insurance firms. So one could analyze insurance PSM letters in a similar way. Second, we used random forests in this article, a type of supervised machine learning, where input data (the letters) were labeled. Other researchers could alternatively use an unsupervised machine learning approach (Chakraborty and Joseph, 2017). This might identify clusters of firm letters quite apart from their category and the regulatory regime under which they were sent. Finally, the analysis could be extended to study supervisory communication across jurisdictions. This could be an especially useful exercise to do, for instance, by looking at letters to firms operating in multiple countries to identify if there are any differences between how 'home' and 'host' regulators communicate with the same financial firms.

Appendix A

A1 Random Forest Variable Importance

Below is an explanation of how the variable importance score of a feature was calculated, using salutation style as an example. Recall the first decision tree from Fig. 2.

The decision schema derived from the training data was then used to classify letters out-of-sample in the test data. As in Fig. 2, the initial accuracy of Tree 1 was 67%, as two out of three letters (Letter 2 and Letter 9) were classified correctly, while one letter was not (Letter 6).

Observation	Features		Predicted	Actual
	Salutation style	Letter length		
Letter 2	Informal	<1,500 words	CAT 1	CAT 1
Letter 6	Formal	>1,500 words	CAT 1	CAT 2–4
Letter 9	Formal	<1,500 words	CAT 2–4	CAT 2–4
Tree accuracy (predicted/actual)				67%

Now, to calculate the variable importance score for salutation style, this feature was permuted, i.e. shuffled in the test data, while all other features such as letter length were held constant. The permuted test sample was then classified according to the decision rules given by Tree 1.

Observation	Features		Predicted	Actual
	Salutation style	Letter length		
Letter 2	Formal	<1,500 words	CAT 2–4	CAT 1
Letter 6	Informal	>1,500 words	CAT 1	CAT 2–4
Letter 9	Formal	<1,500 words	CAT 2-4	CAT 2–4
Tree accuracy (predicted/actual)				33%

As a result of the permutation of salutation style, only one letter (Letter 9) was classified correctly by the tree. The model accuracy fell to 33%. Thus, the decrease in accuracy was $0.67 - 0.33 = 34\%$.

The variable importance for salutation style was then calculated as the average of the decreases in model accuracy when salutation style was permuted, holding all other features constant, applying each decision schema from all the trees in the forest.

This calculation was then made for the other 10 linguistic features.

Appendix B

B1 *C*-statistic

In this section, we clarify the computation of the *C*-statistic with a toy example. Suppose that we used our model to predict the probability of a letter being a PRA letter based on the letter's features for the six letters in the Table B1. The first column gives the observed letter class (PRA versus FSA) and the second column gives the model's predicted probability of the letter being a PRA letter.

We take all possible pairs of observational units where the first element of each pair is the predicted probability of a PRA outcome for a PRA observation and the second element of each pair is the predicted probability of a PRA outcome for an observation. With these we compute a score (Table B2).

If $\widehat{P}(\text{PRA})_{\text{PRA letter}} > \widehat{P}(\text{PRA})_{\text{FSA letter}}$, the pair is given a score of 1 to indicate that it is 'concordant'.

Table B1 Example of predicted probabilities of a letter being classified as a PRA letter for six letters

Observed	$\widehat{P}$(PRA)
PRA	0.8
PRA	0.7
PRA	0.4
FSA	0.6
FSA	0.5
FSA	0.4

Notes: The first row means that this PRA letter had an 80% probability of being a PRA letter given its features and the random forest model. The last row means that this FSA letter had a 40% probability of being a PRA letter given its features and the random forest model.

Table B2 All nine possible pairs of observations from Table B1 (first and second column) and a flag in the third column for whether the probability in the first column is greater than (= 1), less than (= 0) or the same as (= 0.5) that in the second column

$\widehat{P}(\text{PRA})_{\text{PRAletter}}$	$\widehat{P}(\text{PRA})_{\text{FSAletter}}$	Score
0.8	0.6	1
0.7	0.6	1
0.4	0.6	0
0.8	0.5	1
0.7	0.5	1
0.4	0.5	0
0.8	0.4	1
0.7	0.4	1
0.4	0.4	0.5
		Total = 6.5

For instance, in the first row of Table B2, the predicted probability of a PRA outcome for the PRA observation is 0.8 and the predicted probability of a PRA outcome for the FSA letter is 0.6. Since 0.8 is greater than 0.6, this pair is 'concordant', so we score it 1.

If $\widehat{P}(\text{PRA})_{\text{PRA letter}} < \widehat{P}(\text{PRA})_{\text{FSA letter}}$, the pair is given a score of 0 to indicate that it is 'discordant'. For instance, in the third row of Table B2, the predicted probability of a PRA outcome for the PRA observation is 0.4 and the predicted probability of a PRA outcome for the FSA letter is 0.6. Since 0.4 is less than 0.6, this pair is 'discordant', and so we score it 0.

If $\widehat{P}(\text{PRA})_{\text{PRA letter}} = \widehat{P}(\text{PRA})_{\text{FSA letter}}$, the pair receives a score of 0.5 to indicate that it is neither 'concordant' nor 'discordant'. For instance, in the final row of Table B2, the predicted probability of a PRA outcome for this PRA observation is 0.4 and the predicted probability of a PRA outcome for this FSA letter is 0.4. Since 0.4 is equal to 0.4, the probabilities are neither concordant nor discordant, so we score it 0.5.

Finally, to compute the *C*-statistic, we then take the average of the scores. In the toy example above, this evaluates to $\frac{\text{sum of scores}}{\text{total pairs}} = \frac{6.5}{9} = 0.72$. Returning to our definition, this would mean that in this toy example, the probability, given the model, that a randomly chosen PRA letter will be assigned a higher predicted probability of being a PRA letter compared with a randomly chosen FSA letter is 0.72.

Acknowledgements

The views expressed in this paper are those of the authors, and not necessarily those of the Bank of England or its committees. We are grateful to Marcus Buckmann and the three anonymous reviewers for useful comments. Thanks are also due to members of our research advisory committee: David Baumslag, Peter Barrett, Ron Baxter, James Buckley, David Eacott, Peter Eckley, Eric Engstrom, Al Firrell, Marcela Hashim, Nick Hulme, Russell Jackson, David Jeacock, Tahir Mahmood, Mike Mitchell, Paul Munday, Simon Milward, Dimitris Papachristou, Misa Tanaka, Matthew Willison and Pawel Zabczyk. We also thank the Bank's Research Steering Committee for their support, especially Sujit Kapadia, Andy Murfin and Paul Robinson. A special thanks to Simon Morley for vetting the article, Alex Clark and Matthew Gill for suggesting this research, Arthur Turrell for his suggestion of the original title and to Angelina Carvalho, Matthew Everitt and Pedro Santos for their early contributions.

References

Andersson, M., Cerps, U., and Noréus, M. (2013). The case for analytical supervision: a Swedish perspective. *Financial Supervision in the 21st Century*. pp. 33–46.

Baayen, R. H., Endresen, A., Janda, L. A., Makarova, A., and Nesset T. (2013). Making choices in Russian: pros and cons of statistical methods for rival forms. *Russian Linguistics*, **37**: 253–91.

Baker, A. (2010). Restraining regulatory capture? Anglo-America, crisis politics and trajectories of change in global financial governance. *International Affairs*, **86**(3): 647–63.

Bank of England (2018). *The Prudential Regulation Authority's Approach to Banking Supervision*. London: Bank of England.

Bholat, D. and Gray, J. E. (2013). Organizational form as a source of systemic risk. *Economics*, **7**: 1–35.

Bholat, D., Broughton, N., Parker, A., Ter Meer, J., and Walczak, E. (2018). Enhancing central bank communications with behavioural insights. *Bank of England Staff Working Papers 750*.

Bholat, D., Broughton, N., Ter Meer, J., and Walczak E. (2018). *Simply is Best: Enhancing Trust and Understanding of Central Banks through Communications*. Bank of England, Bank Underground.

Biber, D. (1991). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Blinder, A. S., Ehrmann, M., Fratzscher, M., De Haan, J., and Jansen, D.-J. (2008). Central bank communication and monetary policy: A survey of theory and evidence. *Journal of Economic Literature*, **46**(4): 910–45.

Bodnaruk, A., Loughran, T., and McDonald, B. (2015). Using 10-k text to gauge financial constraints. *Journal of Financial and Quantitative Analysis*, **50**(4): 623–46.

Bozanic, Z., Roulstone, D. T., and Buskirk, A. V. (2018). Management earnings forecasts and other forward-looking statements. *Journal of Accounting and Economics*, **65**(1): 1–20.

Brand, C., Buncic, D., and Turunen J. (2010). The impact of ECB monetary policy decisions and communication on the yield curve. *Journal of the European Economic Association*, **8**(6): 1266–98.

Breckenridge, J., Farquharson, J., and Hendon R. (2014). The role of business model analysis in the super vision of insurers. *Bank of England Quarterly Bulletin*, **2014**: 49–57.

Buncefield Major Incident Investigation Board (2012). *The Report of the Buncefield Major Incident Investigation Board into the Policy and Procedures of the Health and Safety Executive's and the Environment Agency's Role in Regulating the Activities on the Buncefield Site under COMAH Regulations*. London.

Capuder, K. M. (2011). Are financial regulators competent? Examining the evidence. *International Proceedings of Economics Development & Research*, **5l**: 99–103.

Chakraborty, C. and Joseph, A. (2017). Machine learning at central banks. *Bank of England Staff Working Papers 674*.

Conrad, C. and Lamla, M. J. (2010). The high-frequency response of the EUR-USD exchange rate to ECB communication. *Journal of Money, Credit and Banking*, **42**(7): 1391–417.

Cutler, D. R., Edwards, T. C., Jr., Beard, K. H., et al. (2007). Random forests for classification in ecology. *Ecology*, **88**(11): 2783–92.

De Nederlandsche Bank (2010). *From Analysis to Action. Action Plan for Change in the Conduct of Supervision*. Amsterdam: De Nederlandsche Bank.

Dent, K., Westwood, B., and Segoviano Basurto, M. (2016). Stress testing of banks: an introduction. *Bank of England Quarterly Bulletin*, **2016**: 130–41.

Dowd, K. and Hutchinson, M. (2014). How should financial markets be regulated. *Cato Journal*, **34**: 353–388.

Dowd, K., Hutchinson, M. O., and Ashby, S. G. (2011). Capital inadequacies: the dismal failure of the Basel regime of bank capital regulation. *Cato Institute Policy Analysis No. 681*.

Duroux, R. and Scornet, E. (2018). Impact of subsampling and tree depth on random forests. *ESAIM: Probability and Statistics*, **22**: 96–128.

Financial Conduct Authority and Prudential Regulation Authority (2015). *The Failure of HBOS Plc (HBOS)*. London: Bank of England.

Financial Services Authority (2009). *The Turner Review: A Regulatory Response to the Global Banking Crisis*. London: Financial Services Authority.

Financial Services Authority (2011). *The Failure of the Royal Bank of Scotland*. London: Financial Services Authority.

Financial Services Authority Internal Audit Division (2008). *The Supervision of Northern Rock: A Lessons Learned Review*. London: Financial Services Authority London.

Garcia, G. G. H. (2009). Ignoring the lessons for effective prudential supervision, failed bank resolution and depositor protection. *Journal of Financial Regulation and Compliance*, **17**(3): 186–209.

Haldane, A. and McMahon, M. (2018). Central bank communications and the general public. *AEA Papers and Proceedings*, **108:** 578–83.

Hansen, S., McMahon, M., and Prat, A. (2017). Transparency and deliberation within the FOMC: a computational linguistics approach. *The Quarterly Journal of Economics*, **133**(2): 801–70.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *Jama*, **247**(18): 2543–46.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC.

Hayo, B., Kutan, A. M., and Neuenkirch M. (2012). Communication matters: US monetary policy and commodity price volatility. *Economics Letters*, **117**(1): 247–9.

Holmes, D. R. (2013). *Economy of Words: Communicative Imperatives in Central Banks*. Chicago: University of Chicago Press.

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied Logistic Regression*. Hoboken, NJ: John Wiley & Sons.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*. New York: Springer.

Jockers, M. L. (2014). *Text Analysis with R for Students of Literature*. Cham: Springer.

Johnson, S. and Kwak, J. (2011). *13 Bankers: The Wall Street Takeover and the Next Financial Meltdown*. New York: Vintage Books.

Jurafsky, D. and Martin, J. H. (2014). *Speech and Language Processing*. Harlow, Essex: Pearson.

Kane, E. J. (2016). A theory of how and why central-bank culture supports predatory risk-taking at megabanks. *Atlantic Economic Journal*, **44**(1): 51–71.

Kellermann, A. J., de Haan, J., and de Vries, F. (2013). *Financial Supervision in the 21st Century*. Berlin: Springer-Verlag.

Kellermann, A. J. and Mosch, R. H. J. (2013). Good supervision and its limits in the post-Lehman era. *Financial Supervision in the 21st Century*. pp. 1–16.

Lastra, R. M. (2013). Defining forward looking, judgement-based supervision. *Journal of Banking Regulation*, **14**(3–4): 221–7.

Li, F. (2010). The information content of forward-looking statements in corporate filings—a naïve Bayesian machine learning approach. *Journal of Accounting Research*, **48**(5): 1049–102.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, **66**(1): 35–65.

Loughran, T. and McDonald, B. (2016). Textual analysis in accounting and finance: a survey. *Journal of Accounting Research*, **54**:(4): 1187–230.

McEnery, T. and Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

National Commission on the BP Deepwater Horizon Oil Spill and Offshore Drilling (2011). *Deepwater: The Gulf Oil Disaster and the Future of Offshore Drilling, Report to the President*. Washington: U.S. G.P.O.

Proudman, J. (2018). Cyborg supervision – the application of advanced analytics in prudential supervision. *Bank of England speech made by James Proudman on 19 November 2018*. https://www.bankofengland.co.uk/-/media/boe/files/speech/2018/cyborg-supervision-speech-by-james-proudman (accessed September 19, 2019).

Ranaldo, A. and Rossi, E. (2010). The reaction of asset markets to Swiss National Bank communication. *Journal of International Money and Finance*, **29**(3): 486–503.

Richards, H. (2013). Influence and incentives in financial institution supervision. *Financial Supervision in the 21st Century*. pp. 73–102.

Sijbrand, J. and Rijsbergen, D. (2013). Managing the quality of financial supervision. In *Financial Supervision in the 21st Century*. pp. 17–32.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics*, **8**(1): 1471–2105.

Strobl, C., Malley, J., and Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, **14**(4): 323–348.

Tagliamonte, S. A. and Baayen, R. H. (2012). Models, forests, and trees of York English: was/were variation as a case study for statistical practice. *Language Variation and Change*, **24**(2): 135–78.

Treeratpituk, P. and Giles, C. L. (2009). Disambiguating authors in academic publications using random forests. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, New York: ACM, pp. 39–48.

Vahlne, N. (2017). On LPG usage in rural Vietnamese households. *Development Engineering*, **2**: 1–11.

Viñals, J. and Fiechter, J. (2010). The making of good supervision: learning to say no. *IMF Staff Position Note.* https://www.imf.org/external/pubs/ft/spn/2010/spn1008.pdf (accessed September 19, 2019).

Notes

[1] Microprudential regulation and supervision refer to rules and their enforcement by regulators, which are targeted at individual institutions. It can be contrasted to macroprudential regulation and supervision that refer to rules and their enforcement by regulators, which are targeted at the financial system as a whole. Microprudential regulation often focuses on the solvency, asset mix, liquidity, business model, and corporate governance of financial institutions. This is different from conduct regulation that focuses on consumer protection.

[2] Although central banks can in theory set negative interest rates on reserve, banks retain the option of withdrawing those reserves and converting them into (zero-yielding) cash. Therefore, zero is the effective lower bound for interest rates.

[3] Multicollinearity refers to a high linear correlation among independent variables. If multicollinearity exists, then the size and sign of coefficients for each variable may be incorrect.

[4] Perfect separation happens when a feature or set of features is completely correlated with the target variable. As a result, the maximum likelihood estimates for these features do not exist and the standard logistic regression will not work.

5 Note that the management of the firm under scrutiny is not involved in this process.

6 We thank an anonymous reviewer for querying this.

7 We are unable to disclose further breakdowns of the letters as a condition of their use for research purposes.

8 Note that as the letters are a collective product whereby no single author can be identified, it was not possible to include a feature to assess the impact of difference in idiosyncratic writing styles. We thank an anonymous reviewer for querying this.

9 We are grateful to Ash Asudeh, Marc Alexander, Tom Bartlett, Andreas Buerki, Billy Clark, Ben Clarke, Rachel Edmonds, Nigel Fabb, Nikolas Gisborne, Andrew Hardie, Christopher Hart, Eleni Kapogianni, James Murphy, and Bonnie Webber for useful comments.

10 https://sraf.nd.edu/textual-analysis/resources/ (accessed September 19, 2019).

11 Part-of-speech tagging assigns lexical classes to each word in a document, such as verb, noun, and adjective.

12 Regularization refers to a set of techniques in machine learning to ensure that models do not overfit to the training data and can generalize well out of sample.

13 We thank an anonymous reviewer for suggesting a penalized logistic regression model as a complementary approach.

14 There are two different ways to build samples for a random forest. The first way is the use subsamples, as we have done. In a subsample, we sample ~63.2% of full dataset without replacement. The second way is to use bootstrap samples. In a bootstrap sample, we draw samples with replacement from the full dataset such that the size of the bootstrap sample is exactly the same size as the original dataset; thus, some observations may occur more than once, others not at all. In a bootstrap sample, it can be shown that ~63.2% of the dataset is included in the bootstrap sample. In practice, we have found that it does not seem to affect performance if subsampling or bootstrapping is done. We prefer subsampling because it is more straightforward to explain pedagogically, but for other reasons why subsampling might be preferred we refer the interested reader to (Strobl *et al.*, 2007; Duroux and Scornet, 2018).

15 Note that for reasons of clarity, we do not show the entire tree as in random forests they are grown deep and it would not fit on the page!

16 To demonstrate the robustness of our approach, as one reviewer requests, we compare our approach with other research papers that have used random forests. For example, Cutler *et al.* (2007) run 5 forests of 50 trees; Treeratpituk and Giles (2009) run 10 forests of 500 trees; in Tagliamonte and Baayen (2012), it appears that a single random forest is grown.

17 Note that *C*-statistic and AUC should be explained differently as they are conceptually different, but are equivalent in value.

18 To validate this result and to diagnose overfitting, we also performed leave-out-one cross-validation. In leave-one-out cross-validation, the dataset is divided into *n*-folds, where *n* is the number of observations in the dataset (in the case of the regulator, dataset 220 letters). We then build the forest on the dataset excluding a given letter, and use the resulting model to predict that letter's class label; this is done for each of the *n* observations. Using this method, classification accuracy and the *C*-statistic are very similar (89.59% and 0.9392, respectively) to those obtained using the out-of-sample data within the random forest. These results combined indicate that our random forest model does not overfit and performs well in classifying letters according to regulator using linguistic features.

19 We followed the bootstrap procedure to derive the standard errors for the confidence intervals, as described in James *et al.* (2013) and—in more detail—Hastie *et al.* (2015). Utilizing the R boot package, we drew $B = 1{,}000$ bootstrap samples with replacement from the original datasets. We built a logistic ridge regression to each bootstrap sample, following the procedure discussed above in Section 4.2, and collected coefficients for each feature. These were stored in a 1,000 × 12-matrix. From this matrix, we computed the standard error of each coefficient as $se_\beta = \sqrt{\mathrm{Var}(\beta)}$, which, together with their means, were used in constructing the confidence intervals. As an anonymous reviewer queries, we do acknowledge that there are problems with constructing robust confidence intervals for penalized models given that the estimates are not unbiased. However, in a relatively exploratory study such as ours that already uses heuristic measures (for instance, the random forest variable importance threshold), we find that these pseudo confidence intervals can give us some indication as to which variables to focus the discussion on.

20 For an additional test of robustness (i.e. to check for overfitting), we also computed performance based on leave-one-out cross-validation: Accuracy = $97.16\%, C = 0.9737$. These are virtually the same, and thus indicate that we can be confident in the quality of the random forest model.